SECURITY CLASSIFICATION OF THIS PAGE

DTIC FILE COPY (1)

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS NONE
2a. SE[illegible] AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.
2b. DI	
4. PE[illegible] AD-A220 855	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFIT/CI/CIA-90-004D
6a. NAME OF PERFORMING ORGANIZATION AFIT STUDENT AT Univ of Texas/Austin 6b. [illegible]ICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION AFIT/CIA
6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code) Wright-Patterson AFB OH 45433-6583
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification) (UNCLASSIFIED)
Utility as a Function of Selection Ration and Base Rate: An Empirical Investigation of Military Aviation Selection

12. PERSONAL AUTHOR(S)
Judy Dennis Roomsburg

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
THESIS/DISSERTATION	FROM _____ TO _____	1990	160

16. SUPPLEMENTARY NOTATION
APPROVED FOR PUBLIC RELEASE IAW AFR 190-1
ERNEST A. HAYGOOD, 1st Lt, USAF
Executive Officer, Civilian Institution Programs

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

DTIC ELECTE APR 25 1990 S D B

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL ERNEST A. HAYGOOD, 1st Lt, USAF	22b. TELEPHONE (Include Area Code) (513) 255-2259 22c. OFFICE SYMBOL AFIT/CI

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

AFIT/CI "OVERPRINT"

UTILITY AS A FUNCTION OF SELECTION RATIO AND BASE RATE: AN EMPIRICAL INVESTIGATION OF MILITARY AVIATION SELECTION

by

JUDY DENNIS ROOMSBURG, B.S., M.A., M.A.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 1989

UTILITY AS A FUNCTION OF SELECTION RATIO AND BASE RATE: AN EMPIRICAL INVESTIGATION OF MILITARY AVIATION SELECTION

Publication No. _____

Judy Dennis Roomsburg, Ph.D.
The University of Texas At Austin, 1989

Supervising Professor: Joseph M. Horn

Research into utility analysis in the context of personnel selection has recently become more prevalent. Difficulties in calculating the utility formula parameters previously hindered their application in the applied setting. An increase in utility analysis awareness occurred, however, with the work of Schmidt, Hunter, McKenzie, Muldrow and their colleagues. The present study applied the theory of utility analysis to the evaluation of the test method of pilot selection in the United States Air Force (USAF) with emphasis on variable selection ratios and base rates.

The Air Force Officer Qualifying Test (AFOQT)

Pilot Composite is one of the main predictors of success in Undergraduate Pilot Training (UPT). Low correlation between the AFOQT and UPT outcome casts possible discredit on its usefulness when taken as the indicator of the value of the AFOQT.

Analyses were performed to determine dollar criterion utility of the AFOQT. Utility gain from use of the AFOQT over random selection was significant for the four year period and sample (N=1550) investigated.

Background information furnished by the subjects upon application to UPT provided supplementary data with which to predict UPT outcome. Through a discriminant analysis process, the background data resulted in more accurate prediction of actual UPT outcome than did the AFOQT alone. Combined, the AFOQT and the background data resulted in an even greater gain in utility than either generated alone.

In addition to the primary emphases, development of a new concept (Delta Base Rate) involving base rate evaluation in the multiple hurdle selection

process was introduced. Delta Base Rate was shown to go beyond the traditional definition of success ratio when considered in a multiple hurdle selection process.

The results of the study were interpreted for the applied setting which demonstrated the benefit of incorporating utility analysis into the personnel selection decision process. Implications for the need of further research in personnel selection utility analysis involving nonprofit organizations were addressed.

TABLE OF CONTENTS

INTRODUCTION

Strict psychometric interpretation of psychological research has resulted in confusion surrounding selection programs in applied settings. The jargon of industrial and personnel psychologists contains references to concepts that are alien to corporate users. Terms such as 'statistical significance', "percentage of variance accounted for", and "confidence intervals" make no sense to the very individuals who demand usable outcomes of research. Confusion also exists even within the psychological community concerning basic definitions used in the process of selecting employees. Divergent meanings associated with terms such as "decision theory", "utility", and 'productivity" result in confusion in the interpretation of research which limits the full applicability of selection instruments.

The traditional emphasis in research by psychologists has been on increasing the validity of

the selection instrument. The higher the predictive validity between the selection device and behavior on the job (criterion), the better the selection instrument. An acceptable "level" of validity has often been the end product of research. Validity coefficients, however, generally have limited meaning tc non-psychologists in industry.

As members of the total management and organizational team, psychologists must learn to equate their findings to what non-psychologists know. They must realize that they too are in business - the business of selection - and their resources are the selection tools used to improve their programs. Industrial psychologists cannot become bonafide members of the management team when, for any given discussion between them and corporate executives, half of the conversants are unable to understand half of the arguments whether the question concerns business or psychology (Darlington, 1976).

This translates to a need for "dollar talk" - the demonstration of the efficiency of a selection device by the description of the criterion (used for validity estimation) in dollar terms. Basic inferential statistics courses now emphasize "practical" significance or the use of common sense in describing research findings. For example, Hays (1981) warns against overemphasizing statistical analysis and "confusing the paintbrush with the painting" (p. 265) when determining which findings are useful and which provide solely statistical significance. Regrettably, reliable methods for determining practical significance are not as readily available as the warnings to use them.

It is unreasonable to believe that management can or should become proficient in psychological research interpretation. Their very purpose for hiring personnel consultants is to gain organizational improvements from an area of expertise outside of their own domain. It is, therefore, incumbent upon the industrial

psychologist to be able to translate his or her findings into management terms. If the term "significant" can be adapted to mean "500,000 dollars' worth of gain in organizational productivity by newly hired employees", then management can understand the research they have been paying for and psychologists will be able to establish a common language for the two sides. Only then will the optimal use of personnel selection research be realized.

The magnitude of the validity coefficient found in personnel selection research has been notoriously low (compared to empirical laboratory studies) despite efforts by psychologists to improve them. Validity coefficients of 0.40 have commonly been accepted in selection studies (Guion, 1965). A validity of unity, possible in some laboratory studies, is unheard of in practical selection research and 0.70 is considered the ceiling value (Brogden, 1949). Efforts to show practical usage

with low correlations resulted in the development of tables incorporating validity values with other selection environment parameters which could be more easily translated into decision-making options (Taylor and Russell, 1939). Criticisms of the Taylor and Russell tables (Smith, 1948; Brogden, 1946, 1949; Schmidt, Hunter, McKenzie, & Muldrow, 1979) are common but their work is important because it established the importance of the base rate and selection ratio parameters in selection decisions.

Personnel research thus progressed from a predictive validity emphasis, in the form of correlation coefficients to predictive accuracy (through the use of decision making theory), in the form of proportion of correct predictions. Personnel selection decision theory is the application of a mathematical model to the selection process. The decisions involve maximizing the number of successful people out of those selected (valid accepts) while minimizing the number of people accepted who will eventually fail (false

accepts). All selection decisions will result in these two outcomes as well as the correct rejection of people who would have failed (valid rejects) and rejection of some people who would have succeeded if given the chance (false rejects).

The translation of research results into tangible organizational outcomes (i.e., dollar figures) has been a natural extension from validation and decision theory studies. Validation demonstrates the statistical significance of the test instrument, decision theory translates the validity into selection alternatives and predictive utility assigns a dollar amount to the various options.

Studies which introduced the monetary concept into selection decisions have existed for over forty years (Brogden 1949; Brogden and Taylor, 1950; and Cronbach and Gleser, 1957) but the results were not implemented operationally due to difficulties in assigning dollar figures to performance differences

among employees (Cascio, 1982). Inability to monetarily quantify performance precluded the determination of the dollar amount assigned to a "standard deviation of performance" (SDy) term required for the utility formulas. The estimation of this SDy value became the "achilles heel" of utility analysis implementation (Cronbach & Gleser, 1957).

In addition, numerous research studies implied the existence of validity specificity (and therefore, utility specificity) which required the revalidation of each selection instrument in each employment setting. The lack of validity generalization and difficulties determining the SDy were predominantly the causes for psychologists discounting utility analysis and focussing on the psychometric demonstrations of the efficiency of selection instruments (Schmidt, Hunter, McKenzie, & Muldrow, 1979).

The work of Schmidt, Hunter, McKenzie, and Muldrow kindled a renewal in the interest of

quantifying performance into economic (i.e. utility) terms. Through studies demonstrating validity generalization, they concluded that validity estimations could be generalized across job families. This finding also opened the door for generalizing utility across comparable jobs as a function of validity. Their analyses, described more fully in the next section, resulted in a relatively stable and replicable estimate of 40 percent of the mean wages (for the position in question) as the standard deviation of performance in dollars. Since the demonstration of the usefulness of this index, utility estimations are now becoming an added extension to traditional selection instrument research.

The continued progression from validity to utility will increase the ability of industrial psychologists to provide better decision alternatives to management and to interpret the results more readily to non-psychologists.

Predictive utility allows the end-user of psychological research to justify current selection programs on a cost-effectiveness basis as well as make more realistic comparisons between alternative selection programs which may be under consideration.

Predictive utility, however, even with recent advances has not reached a high level of acceptance due to an excessive reliance on simulations and "make-believe situations" (Landy, 1986). He stated that a "simple technology rather than theory building, verification or extension" was represented in previous research. Thus, future utility analyses must focus on establishing theoretical and more importantly, real-life bases.

Empirical utility analyses involving real organizations have only recently emerged (Arnold, Rauschenberger, Soubel, & Guion, 1982; Schmidt, Mack, & Hunter, 1984; Hogan & Zenke; 1986; Schmidt, Hunter, Outerbridge, & Trattner, 1986; and Schmidt, Hunter, & Dunn, 1987). The importance of the justification through utility analyses for continued

use of a current selection instrument has not been a major focus of the research to date. It is important to examine existing selection instruments in real organizations not only to justify the current selection procedures but to provide a baseline for future utility analyses of potential selection instruments.

In addition, information on the impact of utility analysis in non-profit organizations and particularly for those jobs which do not have a tangible outcome (i.e. a production unit) is lacking (Schmidt and Rauschenberger, 1986). Particularly evident is the failure to emphasize the actual operational variability of the equation parameters (i.e. selection ratios, base rates - or base rates of success for different steps in a multiple hurdle process - and validity coefficients). The lack of empirical demonstrations of the utility formulas in these situations has resulted in a gap in the industrial and personnel psychology literature.

The main emphasis for the present study concerns selection ratios, base rates of success in training and the investigation of their variable impact on utility in the United States Air Force (USAF) pilot selection program. Although it would be unrealistic to assume (particularly in large organizations) that exact selection ratios and base rates of success can be derived, very close approximations, using true current and historical organizational personnel data can be achieved. This would permit the calculation of actual utilities that result from the different combinations of parameter values. Boudreau (1989) has addressed this type of investigation under the rubric of "sensitivity analysis". Sensitivity analysis is the process where all but one of the parameter values are held constant. The lowest to highest values of the variable parameter are examined in different combinations with the constant parameters to see which parameter variation has the largest impact on utility. The present study departs from sensitivity

analysis in that it will investigate simultaneous changes in all of the parameters.

The present research investigates the implementation of utility analysis in the USAF pilot selection program. This study is intended to demonstrate empirically, that utility analysis in the context of military aviation training selection procedures is not only feasible but informative. Through the incorporation and extension of past theory and methodology concerning utility analysis, one goal will be to demonstrate the benefit of incorporating utility analysis in the justification for continued use of the Air Force Officer Qualification Test (AFOQT). The AFOQT has continued to be used for selection of military aviators despite its low correlation with performance in undergraduate pilot training (UPT).

A second goal will be to investigate incremental validity and the increase in utility that follows the addition of background information.

Early versions of the AFOQT included a biographical inventory; however, for various reasons this biographical portion was eliminated.

Only two studies investigating the AFOQT were found in a comprehensive literature review of utility analysis. One involved a life cycle costs approach which estimated, in part, the cost of AFOQT administration. Included component costs were research and development, acquisition, and operation and support (Bortner & Ree, 1977). The life cycle costs approach considers historical costs and price quotes to derive the component estimates. A benefit of the Bortner and Ree research for the present study results from their development of estimates of a five year period which, when averaged, provides an empirical derivation of the actual cost of testing per applicant.

The other reference provided a brief description of a comparison of three utility strategies involving a one-cohort analysis (Roach, 1983). The lack of detailed information in the

report does not allow for replication or verification of the study. Utility analysis of background data in the context of military aviation training selection has not been performed.

A final consideration will be to contrast the Schmidt, Hunter, McKenzie, and Muldrow (1979) version of utility analysis with that of a strict cost-benefit analysis modified from the McCollom-Savard Direct Method described by McCormick and Tiffin (1974). This method consists of dividing applicants into deciles based on standardized percentile scores and determining pass/fail rates within each decile. Actual failure (attrition) rates can be determined for any given cutting score on the predictor and cost-benefit determination can be made. The Direct Method is, in essence, the current method of establishing the usefulness of any selection method in the USAF.

The present study will utilize the personnel database currently maintained by the USAF Human

Resources Laboratory. Samples of pilot candidates who entered UPT over a four year period will be obtained as well as their AFOQT scores and background data items. The selection ratios and base rates of success in training will be calculated for each year group. Correlation coefficients between AFOQT scores and the performance criterion of UPT outcome can also be obtained for each cohort. These terms will be entered into utility formulas suggested by Schmidt, Hunter, McKenzie and Muldrow (1979) to derive an estimated dollar value from the use of the AFOQT for each year group.

Predictor items from various background data available in the database will then be analyzed through a discriminant analysis classification procedure with UPT graduation or elimination as the grouping variable. The incremental validity and utility of the addition of these background data items to the present selection system will then be assessed.

Before addressing the design in greater detail,

a review of the literature and previous research findings on military pilot selection, background data and utility analysis will be discussed. These findings and the establishment of the importance of utility analysis in the determination of the usefulness of the AFOQT and the inclusion of background data in the prediction of performance will yield the hypotheses of the present research.

CHAPTER I

LITERATURE REVIEW AND BACKGROUND

Fiscal budget constraints and rising costs of training have increased the necessity for cost effective aviator selection in the Department of Defense (Kantor & Bordelon, 1985). Failures in undergraduate pilot training (UPT) result in loss of time and money to the United States Air Force (USAF). In 1985, average loss of USAF investment in each person who eliminated from UPT was determined to be $67,000 (Kantor & Carretta, 1988). This loss in investment, though substantial, does not reflect the "true" actual expense to the government. The benefits that would have been realized by selecting and training an individual who would have succeeded in the eliminated person's place must also be considered. It is obviously better to select those who have a better than average chance of succeeding in the on-the-job (operational) performance criterion.

The actual criterion of operational performance would be used to select employees if time and funding allowed (Brogden, 1946a). In other words, every applicant would be hired and given the chance to perform the job. Those who failed would be discharged and those who could perform the work would continue employment. Since this is generally not feasible, some selection instrument is used, which when validated, closely predicts success on the criterion. If the test correlates perfectly with the criterion (r = 1.00), comparability with using the actual criterion is achieved and selection funds are more beneficially used since only those who will succeed will be hired.

Competition for defense dollars necessitates the continued justification of present selection methods. It is no longer sufficient to simply show that selection tests are statistically valid and reliable. Moreover, in a multiple hurdle training selection program, where applicants compete at successive steps for continuation in the process,

the resultant smaller group becomes progressively more homogeneous on those factors which are requisite for success in training. This restriction in range of ability on these factors reduces the correlation coefficient - the very measure that has traditionally established the worth of a test.

The Air Force Officer Qualifying Test (AFOQT) is the primary predictor currently used and has typically resulted in correlations of only 0.16 to 0.22 with success in UPT (Roach, 1983). These low correlations have been attributed to the consequences of the considerable homogeneity which exists in the selected groups. Regardless of the explanations, or actual reasons, it is difficult to justify the use of the AFOQT based upon its validity with UPT alone. Even correlations derived from restriction in range correction procedures do not accurately reflect the actual usefulness of the test.

Some managers have a basic understanding of the

statistical properties used in test validation and have a true interest in the level of validity established between the predictor and the criterion. For them, additional justification must accompany the validity coefficient. The more typical situation is one in which managers have no understanding of the statistical concepts involved. The solution to this problem of presenting an accurate estimation of the usefulness of a test is to shift the focus from validity to utility. Substantiating the utility as well as the validity of a test is now a major focus of contemporary personnel selection (Greer & Cascio, 1987).

The lack of a common language between those who do the research and those who make decisions based on the research, contributes to the difficulty in demonstrating the utility of a test. It is incumbent upon the researchers to be able to translate their findings into arguments for (or against) implementation in terms that decision makers can comprehend. Translating research

findings into some sort of economic index is intuitively logical (Brogden, 1949) since the majority of organizations depend on dollars either for funding or profit.

A discussion of the historical progression of utility analysis will provide a background for the present study. Some of the major issues of utility analysis development will be addressed and the relationship between the concepts will be considered.

The development of personnel selection utility analysis progressed from the Taylor and Russell (1939) Classic Validity Model (Campbell, 1983) to the Brogden (1949) and Cronbach & Gleser (1965) Classic Utility Model (Alexander & Barrick), 1987) and then on to the Schmidt, Hunter, McKenzie and Muldrow (1979) Global Estimation Utility Model (Greer & Cascio, 1987). This progression began with the realization of the inadequacy and limitations of the validity coefficient when used as the sole

determiner of the benefit of a selection instrument.

The Classic Validity Model

Taylor and Russell (1939) demonstrated that the magnitude of the validity coefficient was only partially indicative of the usefulness or utility of a valid selection instrument. They introduced the relationship between base rate (the proportion of present successful employees), the selection ratio (the number of selectees to the total number of applicants) and the validity coefficient (derived from the selection instrument and performance criterion in question). Their study proposed that the utility of the selection instrument is a function of all three of these measures and increases in base rate (i.e. more of the selected employees being successful) due to the selection instrument can be determined.

Taylor and Russell modified the personnel selection process from a purely statistical method to one which employed a decision-making process by

focussing on the proportion of correct predictions rather than the validity coefficient alone. The demonstration that even a predictor of low validity can benefit the organization if the goal is to select only the best of the applicants (i.e. a very low selection ratio) dispelled the opinion that low validity coefficients were evidence of deficient selection instruments. Figure 1 is a schematic diagram provided by Taylor and Russell for combining validity (represented by the ellipse) with base rate and the selection ratio.

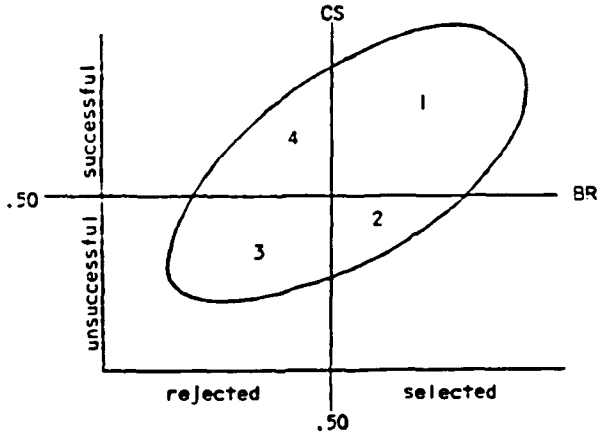

Figure 1. Decision Schematic

Line BR represents the cutting score on the performance criterion. All persons who are currently satisfactory are in the areas designated as 1 and 4. Those that are unsatisfactory are represented as areas 2 and 3. Line CS is the critical cutting score on the selection instrument. The assumption is if the present group of employees had been tested on the test being validated, the prediction would be that those people in areas 1 and 2 would have been selected while those in areas 3 and 4 would have been rejected.

The numbered areas in Figure 1 represent possible selection decisions. Selection of people in areas 1 and 3 would be correct decisions (valid accepts and valid rejects) while people in areas 2 and 4 would be incorrect decisions (false accepts and false rejects).

Valid accepts (area 1 of Figure 1) are those people who score above the critical cutting score on the predictor and who would be successful on the job. Valid rejects (area 3 of Figure 1) are those

who would not be selected by virtue of their test scores and would have failed in required performance. False rejects (area 4 of Figure 1) are those who would not have been selected based on their test scores but would have performed successfully if given the chance. False accepts (area 2 of Figure 1) are those people who score passably on the test but will not perform at the level required to fulfill the job requirements.

The Classic Utility Model

Despite the Taylor and Russell findings, researchers continued to justify the use of a straightforward index of test usefulness based on derivations of the correlation coefficient such as Hull's index of forecasting efficiency (E), Kelly's coefficient of alienation (k) and the coefficient of determination (r^2). The proponents of these correlation alternatives were attempting to equate applied setting correlations with laboratory

findings. However, in doing so they failed to consider the unique parameters found in the applied selection test situation (Schmidt, Hunter, McKenzie, and Muldrow, 1979).

Brogden (1946a) criticized the use of all of these measures in the appraisal of test efficiency. The use of these measures resulted in discounting selection tests where validity with the criterion was less than 0.50. Because tests found in practical situations were frequently less than this, many potentially good instruments were rejected since they could not attain these levels of validity.

Brogden believed that use of correlation derivatives was inappropriate since their main focus was in measurement of standard errors of estimate and not in the direct measurement of test efficiency. He concluded that the main goal of selection prediction was to increase the ability level of the selected individuals (improve the "goodness" of those selected) which involves

determining the variation in the mean criterion score ($\bar{Z}y$) not the error estimate derived from the use of a selection instrument. He determined that the "goodness" of the selected group was a function of their test scores and the magnitude of the product-moment correlation. Since the main point in all of these derivations is errors of prediction ($Zy - \bar{Z}y$) they are only indirectly concerned with variations of $\bar{Z}y$. Neither $\underline{E}$ nor $\underline{k}$ provides a function with mean test scores to obtain an equation which demonstrates the goodness of a selected group.

Brogden (1946a), like Taylor and Russell advocated using the product-moment correlation as a direct measure of the usefulness of a test but disagreed in part with their dichotomous split on the criterion. He stated that due to the curvilinear relationship with the predictor which exists in the upper portion cases of the criterion split (i.e. the satisfactory employees), "either/or" decisions were deemed appropriate only when

decisions by individuals (such as career choice) were made. Information concerning variation in the performance of individuals above the base rate line was also lost when a dichotomous criterion was applied.

Brogden believed that if a test was in fact a measure of the criterion of interest, and a linear relationship existed between the two, the correct index of efficiency was the product-moment correlation which demonstrated the ratio of the mean test score of the selected group to the actual mean criterion score in the population of applicants. In addition, contrary to the Taylor and Russell model, Brogden asserted that the linearity between the predictor and criterion existed (despite different values of the selection ratio) when the frequency distributions of both predictor and criterion were in agreement. Accordingly, he developed a utility formula which provided an estimate of the proportion of maximum saving achieved from the use of a selection instrument.

Brogden defined maximum saving as that realized from using the true criterion as a selector. (Selection on the criterion itself, would in effect, be comparable to using a test with a validity coefficient of unity). The upper limit of validity would be the perfect criterion (1.00), while the lower boundary (disregarding negative correlations) would be the criterion mean of the population (Curtis & Alf, 1969). The ratio of the selected group criterion mean to the mean on the actual criterion would demonstrate the increase in test validity actually attained.

A realistic measure of test efficiency would demonstrate the proportion of this savings realized from the use of a valid test as denoted by the correlation coefficient. A validity of zero would equal the same efficiency found in random selection since even random selection results in some efficiency (i.e. some selectees perform adequately). Brogden described a method for deriving a product

moment correlation that reflected this ratio of the increase gained from selecting above a critical cutting score on the predictor to the increase that would be obtained from the same critical score on the actual criterion.

Brogden developed another important contribution to utility analysis when he noted that even when defining the product-moment correlation as a ratio of savings, the resultant coefficients would be less than unity when the criterion and predictor distributions *did not agree.* He proposed altering the criterion to construct an equal interval unit distribution based, in effect, upon standard deviations in performance (1946b). The efficiency of the predictor would then be based upon the magnitude of the correlation coefficient which would reflect the association between test scores and the continuous equal-interval criterion.

An equal-interval criterion would allow comparability across units of the criterion where individuals could be compared within a job category,

generalizability between like jobs and comparisons between different organizations. This type of measure would also allow the determination of the importance of a particular job to the overall efficiency of the organization. By using comparable unit measures, such as dollars, a criterion scale could be developed that would have the same meaning at all points of the scale and using a direct measure of the mean criterion score of the selected group ($\overline{Y}$) would reflect the amount saved using a particular selection process.

Brogden formally incorporated a cost of testing factor into his utility formula in 1949 and established what has become the "landmark" in utility model development (Schmidt, Hunter, McKenzie, & Muldrow, 1979). He also emphasized that although there are great increases in organizational efficiency when low selection ratios can be employed, the cost of testing may negate the gains. Cost of testing referred to actual costs of

alternative testing programs. He believed this to be a translation of the test validation process into "cost-accounting terms".

Brogden believed that this cost-accounting approach would also result in the translation of research into practical terms to demonstrate the maximum savings or minimum costs involved in the selection process by stating the criterion in dollar values and formally including a cost of testing variable. Brogden determined the product of the validity coefficient (between predictor and criterion) and the standard deviation of performance on the criterion approximated the mean dollar saving by each unit increase in the standard test score.

The mean of the test score of the selected applicants was assumed to be normally distributed therefore it could be calculated from the formula z/p where "z" represents the height of the ordinate of the normal curve at the critical cutting score on the test and "p" represents the selection ratio. The critical cutting score, in effect, defines the

selection ratio (all persons above the cutting score to all applicants).

The cost of testing was shown to affect the Taylor and Russell assumption that a selection instrument with low validity was still useful if the selection ratio was low. When the cost of testing is high, a low selection ratio can produce prohibitive testing costs. Although the testing cost of a single applicant remains constant, the cost per individual tested fluctuates with changes in the selection ratio (1949). The cost of testing a large number of applicants to select only a few may exceed the cost of the savings realized from the selection instrument. Low validity coupled with low selection ratio, under these conditions, possibly renders a test as deficient as believed prior to the Taylor and Russell findings.

For example, a situation where only one out of a hundred applicants are selected and the cost of testing each one is ten dollars results in 900 of

the 1000 testing dollars being spent to reject applicants. If on the other hand, 90 were selected from that same one hundred, only 100 dollars would be spent to reject applicants. This situation provides evidence that the increase in cost of testing is an inverse function of p.

Brogden suggested that this occurrence results only when the cost of testing is calculated for each applicant. In those circumstances when the cost of the test remains constant regard less of the number tested (for example *in* group testing), the ramifications of testing costs may not be as great.

Brogden and Taylor (1950) first suggested using subjective (expert) judgment as opposed to ratings or other raw or standard score form criteria in determining the weighting of the importance of a particular job within an organization (i.e., the determination of SDy). This advice was the foundation of the decision that resulted in the current definition of SDy. It is also interesting to note that in Brogden's 1949 work, the exact

value of the standard deviation of criterion performance was considered "unimportant and assumed to be unity" (p. 179).

Brogden and Taylor also provided more demonstrations of the value of estimating the criterion in cost-accounting terms. They stated that using dollar interpretations of the criterion allowed for face validity and presumably managerial acceptance of research findings. In addition, they provided rationale for the advantages of the cost-accounting approach. Their interpretation of the cost accounting method provided for the establishment of a "common metric" (i.e. dollars) which permitted the estimation of the interrelations of all important concepts employed in utility analysis, i.e. the validity coefficient, the SDy, the selection ratio and the cost.

Cronbach and Gleser (1957, 1965) extended the concept of utility to different selection strategies as well as classification and placement decisions.

Their most significant and useful contribution was the clarification of the assumptions underlying utility analysis.

Often the goal of utility analysis is to show improvement in selection "over chance". They determined, that in most applied selection programs, some selection or pre-screening has occurred and that chance (or random selection) is generally not representative of the real selection situation. Therefore, they base their work on the assumption of an a priori strategy which defines the population (and therefore the selection ratio determination) at the point where the new instrument is applied.

They also emphasized that the validity coefficient used in utility analysis is not the coefficient obtained from an unselected population but is that obtained from the "a priori population" (the group already pre-screened) being investigated. They go so far as to state that if a zero-order correlation coefficient from an unselected population is used, it must be reduced to be used in

their utility formulas.

Even with the Cronbach and Gleser interpretation, difficulty in assessing the SDy term in the utility formulas resulted in applied psychologists not attempting its practical application. Validity generalization studies beginning in the late 1970's opened the door for a renewed emphasis on utility estimation.

The Global Estimation Model

Current investigation into personnel utility analysis was advanced by research conducted by Schmidt, Hunter, McKenzie, & Muldrow (1979). They demonstrated a method for determining a reasonable dollar estimate of an employee's contribution to an organization. Inability to assign a dollar amount to different levels of performance was the major reason for utility analysis to lay dormant for almost forty years. Prior to the Schmidt, Hunter, McKenzie, and Muldrow (1979) research, assigning

values to variations in performance was the "achille's heel" of utility analyses (Cronbach & Gleser, 1957).

Schmidt, Hunter, McKenzie, and Muldrow (1979) report a pilot study which employed budget analyst supervisors as expert judges to estimate the value of individual productivity, in dollars, of the average analyst's performance. The judges were then asked to determine the difference between the average performance and those analysts who were performing at a level of 85 percent. The difference between the two estimations was then taken as a measure of the worth of one standard deviation of performance.

An extension of this study was conducted to address the three main concerns in utility analysis; the magnitude of utility, a demonstration of utility equations and to test whether the dollar criterion was normally distributed. Evidence was provided for all three areas and substantial utility estimations were presented.

The results of further studies, employing basically the same design demonstrated that the productivity output standard deviation was judged to be 20 percent of the mean output of the group (Schmidt & Hunter, 1981). Other studies have demonstrated that the standard deviation of performance in dollars can be adjudged as 40 percent of the wages of the average person in the job when true calculations of output performance is unobtainable (Schmidt & Hunter, 1983). Use of the 40 percent figure will always give a conservative estimate which will help compensate for any inaccuracy between jobs and of other values used in the utility equations. The 40 percent estimate has been determined to be particularly applicable to training situations (Schmidt & Hunter, 1983).

The Schmidt, Hunter, McKenzie, and Muldrow (1979) work was an extension of both the Classic Validity and Utility Models. The emphasis was on the determination of "marginal" utility as the

"increase in dollar value of average performance that results from using the test" (p. 611). The measure of marginal utility (assuming normally distributed test scores) incorporating costs is:

$$\Delta\bar{U}/\text{selectee} = r_{xy}\ SDy\ \Phi/p - C/p$$

Where:

$\Delta\bar{U}$/selectee = gain in utility from use of test
r_{xy} = validity coefficient between predictor-criterion
SDy = standard deviation of performance in dollars
Φ = the ordinate height of the normal curve at p
C = cost of testing one applicant
p = selection ratio

The usefulness of the selection test is comparable between organizations regardless of the number of selectees since the increase in successful persons selected (increase in base rate) is proportionate to the number of people selected. Since utility is not simply a function of the total number of persons selected, it is comparable across organizations regardless of the actual organizational size.

Selection can be considered as a classification

function if one considers that it involves classifying an individual into either a select or reject category. Since there is a reject category, correct identification and rejection of poor prospects (increase in valid rejects) can result in a substantial utility gain when rejection has a value (Schmidt, Hunter, & Dunn, 1987). This is contrary to the Cronbach and Gleser (1957) position that a rejected applicant has a value of zero in a fixed- treatment selection strategy. Assignment of a dollar value to pilot candidates who are eliminated from UPT (rejects) will be incorporated into relevant analyses of the present study to yield a closer approximation to true utility of the AFOQT.

The Present Model

Utility analysis is the assessment of the economic impact of organizational programs (Katzell & Guzzo, 1983). This impact can be determined through a straight-forward analysis patterned after the Schmidt, Hunter, McKenzie, & Muldrow (1979)

methods, using dollar values. Based on the values calculated, the actual dollar value of the increase in base rate from the use of the selection instrument can be determined. Any evaluation of the economic impact of a selection program must begin with the assessment of the existing selection procedures.

Currently, the USAF selects people to enter UPT based upon a multiple hurdle/multi-stage process with primary emphasis on the Pilot Composite of the AFOQT. In a multiple hurdle selection process, decisions (one of which can be to reject a portion of applicants) are made at each step of the processing. This results in a decreasing number of applicants remaining for the next step. In this applied setting, it is advantageous to decrease the number of eventual trainees due to the considerable costs involved in actual UPT.

Although rejection at each step will generally result in the pilot applicant being used in some

other job within the Air Force, the focus of the present study is on the utility of the present pilot selection program not the USAF personnel selection program in general. Because of this restriction, a departure will be made from the Cronbach and Gleser (1957) assumption that a reject category is assigned a value of zero. In a sub-analysis of the present study, the decision to reject will actually have a value in that once a person is selected and entered into training, another person is not placed into that training slot if the individual fails (regardless of time of elimination). Therefore, once training has started and the individual eliminates, the cost of training one individual is deducted from the utility of the selection instrument in question. This is based on the assumption that with a test of high utility, the selectee would have succeeded and the training cost would have been beneficially expended.

A dipiction of the major hurdles in the current pilot selection program is provided in Figure 2

which will also aid in the description of the current USAF multi-stage selection process.

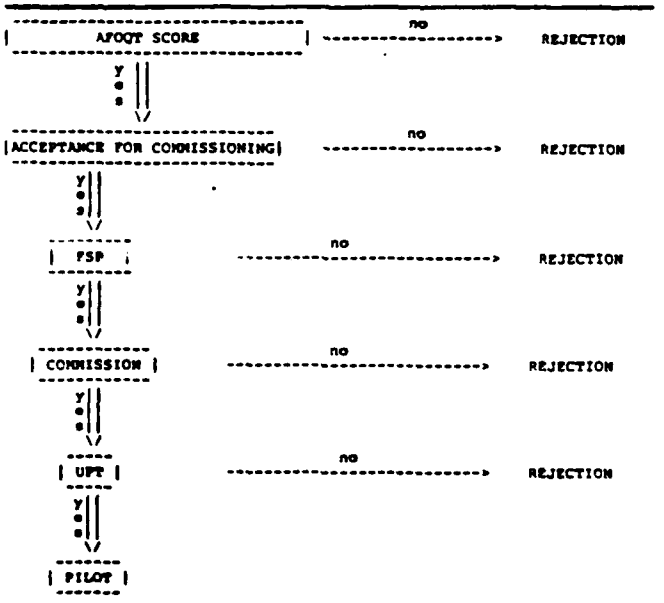

Figure 2. Progression of OTS Candidates in the UPT Selection Process.

Currently the most valid and reliable predictor of success in UPT is cognitive ability and aptitude. Support for the continued use of aptitude has been provided by a meta-analysis conducted by Hunter & Hunter (1984). They found that cognitive ability is the highest predictor for most employment situations.

The AFOQT is the primary test instrument to measure aptitude in the selection of pilot candidates. Revalidation of the AFOQT is periodically accomplished through comparisons with performance in UPT (Miller, 1966, Bordelon & Kantor, 1986). The preliminary version of the AFOQT was originally administered in 1951 (Rogers, Roach, & Short, 1986) with Form A developed in 1953. The primary purpose of its development and current use is to select civilian pilot applicants for precommissioning programs as well as aircrew applicant classification.

Form O is the operational form currently used and was implemented in September of 1981 (Skinner & Ree, 1987). All subjects in the present study were tested on Form O.

The AFOQT is a paper and pencil multiple aptitude battery made up of 16 subtests from which five composites; Navigator-Technical, Academic Aptitude, Verbal, Quantitative and Pilot are

derived. The Pilot Composite has been developed to encompass those areas of aptitude which have been found to best predict success in UPT (Skinner & Ree, 1987). The following eight of the 16 subtests comprise the Pilot Composite:

Verbal Analogies
Mechanical Comprehension
Electrical Maze
Scale Reading
Instrument Comprehension
Block Counting
Table Reading
Aviation Information

If desired, anyone may take the AFOQT, however, it is usually taken only by those individuals who show an active interest in a military commissioning program. Although all examinees take the Pilot Composite of the AFOQT only those who specifically volunteer for pilot training and Officer Training School (OTS) are of interest in the present study.

The USAF OTS is a three-month long post college- graduate school. OTS applicants generally take the AFOQT as they are completing the final requirements for graduation from an accredited

four-year college or are already college graduates. OTS candidates report early to attend the Flight Screening Program (FSP) course conducted at Hondo, Texas, near the OTS site. FSP provides instruction in the basics of flight including 16 actual flying training days in the (training) T-41 aircraft (comparable to the civilian Cessna 172).

The AFOQT is regularly given as a primary step in the multi-phase process. If the individual scores above the 25th percentile on the Pilot Composite (the minimum critical cutting score) he/she proceeds to the next level of processing.

Initial screening to become a pilot includes the basic commissioning requirements of a rigorous physical and high moral standards. Since the mandatory draft no longer exists, self-selection is the primary motivation for application although active military recruitment plays a large part in disseminating information about the possibility of a military career. This has the effect of making more people aware of the career potential than may have

otherwise.

The present study will employ the "fixed-treatment" selection process in that one and only one "treatment" (i.e. pilot training) will be the outcome from selection. Fixed-treatment has various assumptions, one being that persons unaccepted for training will be rejected from the institution (Cronbach & Gleser, 1957). It is assumed for the purposes of the present study, the individual will have no further contact with the "institution" i.e. the pilot selection group.

Another assumption is that training cannot be adapted for each individual based on unique aptitude. Fixed-treatment also results in considerable benefit to the organization even with small increases in validity (Cronbach and Gleser, 1957; Hunter & Hunter, 1984). Due to the extreme restriction in range (and resultant decreased validity coefficients) at any level of the USAF pilot selection process it is assumed that the fixed-treatment assumption will yield a better

picture of utility.

A criticism of the Taylor and Russell (Schmidt, Hunter, McKenzie, & Muldrow, 1979) work concerns the decision to reduce the employee group to a dichotomous grouping of satisfactory or unsatisfactory. For the purposes of the present study, a dichotomous grouping is not unrealistic. While it is recognized that there are generally differences in the level of performance of pilot trainees, those who score above the base rate line graduate and those performing below the base rate line are eliminated.

Another criticism of the Taylor and Russell model is that it deals with only one kind of error (false accepts) and only one kind of correct predictions (high hits). Difficulty is encountered when attempts are made to apply the Classic Validity Model to an actual organization (as opposed to the "ideal" selection model situation).

In the actual rather than the ideal selection

process, two of the possible categories of applicants (false rejects and valid rejects) can only be estimated. The criticism is of a real phenomenon which exists in the actual selection environment. Since generally the only group available has been previously pre-screened, the original group (which contains the false rejects and the valid rejects) will be unavailable to the researcher.

Ordinarily no organization would accept all applicants to see which reject decisions are actually valid and which are false. Parenthetically, a study of the effects of the Flight Screening Program on UPT came close to attaining this achievement (Stoker, Hunter, Kantor, Quebe, & Siem, 1979). A method for estimating the complete validity model will be demonstrated in the use of biodata predictors.

A representation of the present a priori applicant population (Cronbach & Gleser, 1957) tested for any given year is assumed to be a

tested for any given year is assumed to be a normally distributed function as diagrammed in Figure 3:

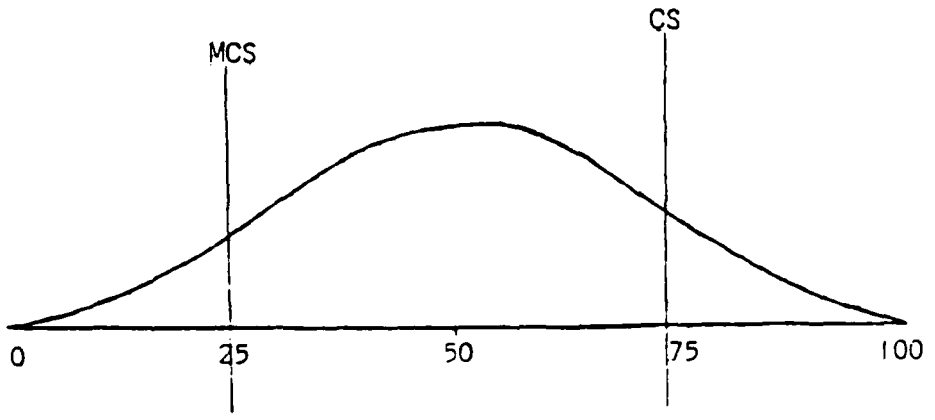

FIGURE 3. A Priori Population

The Pilot Composite scores for a given year (and the a priori population) are charted on the baseline while the vertical line CS represents the critical cutting score on the test for the given year. The proportion of people to the right of the CS line to the total number of applicants is defined

as the selection ratio for that year. Line MCS represents the critical minimum cutting score (Brogden, 1949) which basically does not change from year to year. The MCS (25th percentile on the Pilot Composite) denotes the point (where it is believed) people fall who generally cannot perform at a level to meet the minimum requirements of UPT. There are instances when people who score below the MCS are considered for duty. One instance can be when other items in the application package compensate for scores obtained below the MCS. Another instance for acceptance of people scoring below the MCS can result from an inadequate labor force. Whenever people are selected from lower than MCS scoring categories the potential loss of aircraft and lives caused by known inability will have to be considered.

The cost of AFOQT testing per applicant will be assumed to be $12.12. This amount was derived from an empirical study by Bortner and Ree (1977). They

employed a life cycle cost study which included cost components such as research and development, support, materials, and wages and travel of test administrators in the development of the AFOQT. This amount, per testee, follows the value of ten dollars suggested by Schmidt, Hunter, McKenzie, and Muldrow (1979). The Schmidt, et. al. study as well as the Bortner and Ree findings suggest that the cost per individual tested is the same whether the testing is individual or in a group setting. This contradicts Brogden's theory that the ramifications of testing cost may not be as great when group testing is used.

The Selection Ratio

It is important at this point to address a concern introduced by Alexander, Barrett, & Doverspike (1983). Although they believe that the denominator of the selection ratio is the number of applicants, they dispute the generally accepted definition of the numerator. They state that since the selection ratio (based on the Taylor and Russell

bivariate normal distribution) characterizes the proportion of acceptable applicants in the hypothesized potential population, the numerator must be all potential applicants in that population who score above a critical cutting score. They further state that in adjusting the selection ratio, the actual number of hirees or applicants is irrelevent and the adjustment refers only to increasing or decreasing the selection variable cutting score.

When the sample is a random sample from the applicant population the selection ratio equals a value called the hiring rate. They define the hiring rate is the percentage of applicants that are actually selected. The hiring rate is dependent upon organizational requirements such as quotas, goals and budgetary constraints whereas the selection ratio is dependent upon factors external to the organization. The main objection by Alexander, Barrett & Doverspike is that the majority

of the time in realistic selection research, the two values are not the same because the samples are generally nonrandom.

A consequence of the substitution of hiring rate for selection ratio in nonrandom samples is found when attempting to estimate utility. They state that overestimation of the utility gain results if the hiring rate underestimates the selection ratio.

Different opinions also exist concerning control over the selection ratio. The majority of researchers state that the selection ratio is under the control of the employer (Taylor & Russell, 1939; McCormick & Tiffin, 1974; Cascio, 1982; Siegel & Lane, 1987). Despite the fact that Schmidt, Hunter, McKenzie, & Muldrow (1979) state that the setting of the selection ratio is not under control of the employer (p. 614) they seem to reverse their position when they say that the applicant pool (from which the selection ratio is derived) can be affected by organizational recruiting efforts.

Anastasi (1976) differentiates between external control and internal control of the selection ratio. External control factors are beyond the power of the organization such as when the laws of supply and demand are in effect. Internal control is achieved when the organization can set the cutting score at the point where there is the greatest differentiation between the success and failure criterion groups. These descriptions closely parallel those of the selection ratio and hiring rate terms of Alexander, Barrett, & Doverspike.

For the present study, the definition of selection ratio will be taken as that presented by Cronbach and Gleser (1957). When using the a priori population, the selection ratio can be considered to be those selected to the total number of applicants. The differences found in the size of the a priori applicant population as a function of year group will be considered as only partially under the influence of the USAF in the respect that during

each of the years, it is assumed recruiting efforts existed but were relatively constant. The number selected will be assumed to be under the direct control of the USAF. In addition, the selection ratios used will be assumed to be derived from random sampling of the a priori population for the years under investigation.

It is postulated that the critical cutting score generally referenced in the ideal selection program is, in reality, impossible to determine prior to implementation of the selection program. Only when the selection is completed (or the number needed to fill the vacancies is reached) can a critical cutting score be identified. The critical cutting score for selection is the score achieved by the lowest scoring applicant selected if all vacancies are filled. Only when there are vacancies left after final selection will there be a possibility that an acceptable cutting score exists that is less than the lowest scoring selectee.

References, therefore, to the increasing or

decreasing of the selection ratio (Cascio, 1982) are, in this case, the determination of what percentage of the current applicants will be allowed to enter pilot training. If more are needed, assuming a top-down ranking selection process, more are allowed to enter training. With this assumption, references to "setting a cutting score" is basically like the tail wagging the dog.

On the other hand, selecting fewer applicants does result in moving the CS (Figure 1) line hypothetically to the right which reduces the proportion of false accepts at the expense of increasing the proportion of false rejects (Campbell, 1983). One outcome of the present research will be to examine the yearly fluctutation in selection ratios.

Comparisons between year groups will determine the stability of the the <u>a priori</u> applicant population. If the number of applicants remains fairly stable across years, differences in selection

ratios will be assumed to result primarily from changes determined internally by the USAF. If there are large differences in the size of the a priori applicant population across years, but the total number of selectees remains relatively constant, the selection ratios will be determined to be primarily controlled by external factors. The a priori population (selection ratio denominator) will be defined as the actual number of applicants to OTS-UPT for each of the years under investigation in the present study. The number selected each year for OTS-UPT will define the numerator of the calculated selection ratios.

The Base Rate

While the selection ratio involves cutting scores on the predictor in the population, the base rate pertains to cutting scores on the criterion in the population. As such, the population base rate is the proportion of people who would have been successful (if given the opportunity) to all of the

people in the population (Alexander, Barrett, & Doverspike, 1983). An example of this concept is given by Schmidt (1974) in discussing the Strong Vocational Interest Blank. Comparisons between those who would have succeeded and those who would have failed can be accomplished by extending this definition to denote two populations, one being the "successful" population and the other being the "failure" population.

Normally, the population base rate is not used in applied settings as frequently as the more traditionally defined term of base rate. The term "base rate" in the applied setting has generally referred exclusively to the proportion of the present employees who are successful on present standards and were not selected by means of the test being evaluated (Taylor & Russell, 1939; McCormick & Tiffin, 1974; Anastasi, 1976; Schmidt, Hunter, McKenzie, & Muldrow, 1979; Cascio, 1982; Muchinsky, 1983; Campbell, 1983; Siegel & Lane, 1987). This definition, then, applies to a subsample (i.e.

present employees) of the populations of interest. For the present study, the term "base rate" will be used interchangeably to mean the population base rate or the base rate of present employee group not selected on the selection instrument under consideration.

Increase over present employee base rate remains the ultimate goal of selection with the results generally measured as a "new base rate" (Taylor and Russell, 1939) or success ratio (Cascio, 1982). Cascio states that the success ratio can provide an operational definition of the usefulness of a new test. If the number of people who are successful is greater with the use of a test, other factors being equal, then justification for that test is achieved.

Although, management usually decides what constitutes successful job performance, in actuality, management often bases this decision on what is unsuccessful job performance. In relation

to Figure 1, the traditional definition of base rate equates to the following proportion (in the population):

$$\frac{4 + 1}{1 + 2 + 3 + 4}$$

As stated earlier, in the actual applied setting, areas 3 and 4 generally cannot be determined since available employees, by definition, constitute Areas 1 and 2 for training and only Area 1 for operational employees (i.e. Area 2 employees are eventually terminated). In the training situation, the criterion (pass/fail) cut-off established by management, defines the base rate of success in training.

Two difficulties are encountered when attempting to calculate the traditionally defined value of base rate. The first concerns the concept of the base rate as a parameter of the original unselected population. This renders the base rate as a hypothetical construct by definition. The

second difficulty involves the ultimate criterion (Thorndike, 1949), upon which the traditionally defined base rate is founded. The ultimate criterion is often ill-defined, generally unmeasureable, and in the context of military aviation selection most likely only determined retrospectively and individually (Roomsburg, 1988).

Considering these difficulties, the traditional description of the base rate may not be appropriate in a selection program which consists of a singular test instrument (or test battery composite score) even if selection is based on an easily defined criterion. When training is the criterion with which the base rate is defined, both of the two difficulties just mentioned can be addressed. The traditionally defined base rate, with expanded definition, results in a realistic measure to estimate and apply in an organization that uses a multiple hurdle selection process (MHSP) with an intangible ultimate criterion.

Delta Base Rate ($\underline{/\backslash}$BR) is a term introduced to

allow for the changes in a priori population base rates found in the MHSP where rejection decisions are made at each hurdle. This extension of the base rate term and its meaning, although not following the letter of the traditional definition does follow the intent of the base rate concept when considering multiple hurdle selection.

As stated previously, the MHSP results in rejection decisions at each hurdle with a /\BR available after each hurdle measuring the proportion of people who are successful and pass to the next hurdle. Subscripts will be used to denote which hurdle the particular /\BR pertains to. For example, $/\backslash BR_{MS}$ might refer to the proportion of successful persons resulting from the medical standards hurdle in the pilot selection process.

With this definition in mind, in many applied situations, the "base rate" has, by default, taken on the extended definition of "base rate of success" of a particular hurdle and for any MHSP, many /\BRs

can be calculated and employed to evaluate selection procedures at each level of selection. Following this reasoning, changes in the base rate are dependent, in addition to the a priori population size and selection ratio, upon which hurdle outcome is under consideration and that hurdle's correlation with the ultimate criterion.

These changes in base rate can be designated as $/\backslash BR_1 \ldots /\backslash BR_n$ where $/\backslash BR_1$ equals the success rate after the first hurdle and $/\backslash BR_n$ equals the base rate of the ultimate criterion (i.e. the last "hurdle" in the selection procedure showing the final operational criterion outcome). Technically, $/\backslash BR_n$ is a measure of the proportion of employees who will be successful on a hypothetical construct, the ultimate criterion and has the value of 1.00 since people who theoretically achieve the ultimate criterion are successful. The $/\backslash BR_n$ can only be approached by progressively higher values of $/\backslash BR_1$ through $/\backslash BR_{n-1}$.

In a MHSP, each successive hurdle theoretically produces potential employees who approach the level of satisfactory performance desired for $/\backslash BR_n$ more closely than the previous step or collectively $/\backslash BR_1 \ldots /\backslash BR_{n-1}$. Since each group entering the hurdle is more restricted in range on the characteristics required for the ultimate criterion, it is more difficult to eliminate individuals and the $/\backslash BR$ is higher at each step until, theoretically, the ultimate criterion $/\backslash BR_n$ of 1.00 is achieved.

In the original population, the restriction in range does not exist and therefore the correlation between the hurdle/predictor will be high but the base rate will be low. In the progression of multiple hurdle selection, the correlation becomes progressively smaller and the $/\backslash BR$s become larger.

This rationale is not intuitively obvious since it can be argued that base rates other than those pertaining to the unselected population are based on a progressively restricted subset of the original

population and not comparable to the original population parameter. A hypothetical selection situation will be developed for the purpose of demonstrating that Delta Base Rate can be related to the a priori population as the traditionally defined base rate term is related to the unselected population.

Consider a population of 1000 people and an ultimate operational criterion consisting of extreme requirements. In addition, only 100 have the ability to succeed on this complex ultimate criterion. This would result in a traditionally defined base rate value of 0.10. It might be noted, that if a perfect prediction were possible and only these 100 people were selected no false accept or false reject selection decision errors would result and an accurate selection ratio forecast would be also 0.10. In this situation, a MHSP would be unreasonable since it is designed to progressively "weed out" false accepts and valid rejects from

earlier stages. Elimination of any of these 100 would result in unnecessary false rejection errors.

However, for the present hypothetical situation, consider the organization which employs the MHSP. The very use of a MHSP generally requires originally selecting more applicants than needed to fill the vacancies or satisfy the selection goals. Estimation of false accept and false reject errors and their incorporation into setting selection goals is an inherent part of any MHSP. Generally, a large number of hurdles in the selection process requires the number selected from the original group to be larger than when only one instrument is used. Therefore in a MHSP, the selection ratio (when compared to the traditionally defined base rate) is artifically inflated.

Returning to the hypothetical selection example, the original selection ratio is 0.30 (i.e. 300/1000 people are selected). After the first hurdle 200 people remain in the group to continue in the selection process. This reflects a ΔBR_1 of

0.67 (200/300). After the second hurdle only 150 people remain which results in a $/\backslash BR_2$ of 0.75 (150/200). After the third hurdle only 120 people remain resulting in a $/\backslash BR_3$ of 0.80 (120/150). The final hurdle is training for the ultimate criterion and results in a $/\backslash BR_{n-1}$ of 0.83 (100/120). The correlations for each hurdle are based upon how well the predictor measured in the particular hurdle correlates with the ultimate criterion. Even though the hurdles effectively weed out those individuals who presumably will not succeed on the criterion, the correlations continue to decrease and improvement over each $/\backslash BR$ becomes progressively more difficult. This hypothetical example shows that the concept of $/\backslash BR$ follows the theory found in the personnel selection literature concerning the decrease of correlations as restriction in range increases as a function of a MHSP. What has not been previously emphasized is the function of increasing base rates found in selection.

Many researchers (Anastasi, 1976; Cascio, 1982; Muchinsky, 1983; Campbell, 1983; Roach, 1983) state that the selection measure in question is most useful in absolute (greater gain in the actual number of newly hired employees who are successful) terms when the present base rate is near 0.50. As the base rate deviates from 0.50 in either direction it becomes more and more difficult to obtain correct selection decisions. With their definition, if the base rate is extreme the validity must be higher and/or the selection ratio lower.

This concept can be extended to the /\BR example. Since the correlations continue to progressively decrease in the MHSP (due to the impossibility of obtaining perfectly independent hurdles) and the population selection ratio is artificially inflated to achieve multiple hurdle selection, the concept of hiring rate versus selection ratio comes into consideration. Rather than redefining the selection ratio as "/\SR" along the same lines as /\BR, the hiring rate is employed.

The population selection ratio is more restricted in definition when utility is evaluated (Alexander, Barrett, & Doverspike, 1983). Although the population selection ratio can be extended to apply to the a priori population (Cronbach & Gleser, 1957), changes in the selection ratio as a function of the number of people "selected" after each hurdle of a MHSP would contaminate the results obtained from the utility formulas. The use of the hiring rate, or more specifically the /\HR, at each level of selection is a more justifiable method of determining the usefulness of a the individual predictors in a MHSP. The concepts of /\BR and /\HR will be applied to the present data more fully in the last chapter.

The true UPT base rate for OTS-UPT entrants averaged over years 1970 through 1981 is 0.7424. This figure was calculated from yearly base rates in UPT of OTS-UPT entrants who took AFOQT Forms K through N. This base rate will be used as the

baseline to determine fluctuations over years involved in the present study. Fluctuations from this rate in the present sample as well as fluctuations between years in the study are of primary interest.

Interest in the increase in base rate achieved with the addition of biodata will also be addressed. /\BRs will be established for the both the Pilot Composite and background data.

The base rate resulting from the addition of biodata will be referred to as $/\BR_{BIO}$. In Taylor and Russell (1939) terms, and in relation to Figure 1, the $/\BR_{BIO}$ for the group is:

$$\frac{1}{1 + 2}$$

Areas 1 and 2 represent the UPT trainees who are successful thus, by definition allowing the full prediction validity model of to be estimated for biodata. The UPT trainees who are correctly identified to succeed based on the introduction of

biodata, should be different from the those predicted to succeed based on Pilot Composite alone. This will result if the hiring rate (i.e. $/\backslash HR_{BIO}$ - those selected with the use of biodata) is reduced and the $/\backslash BR_{BIO}$ exceeds the $/\backslash BR_{PC}$ (Pilot Composite).

Cascio (1982) described the presently defined individual $/\backslash$BRs as a measure of the worth of a selection device which demonstrates validity above that found with the present selection method. This increase in validity resulting from the addition of a new selection instrument is called incremental validity.

Use of a particular selection measure must result in more correct selection decisions than would occur without the measure or by chance alone. In other words more "hits" than "misses" would occur in the selection decision process with the use of the new selection device. The present study is developed to demonstrate this increase in "hit"

rate.

One reason to believe that the AFOQT and its low validity still reflects substantial utility is that it is generally agreed that if the predictor has any validity greater than zero, there will be some increase in the base rate (Brogden, 1946, Campbell, 1983). Even if the validity is zero, the number of successful selectees can still be improved (and utility demonstrated) with a lower selection ratio. Even when the selection ratio is not extremely low, if the job in question involves great training costs, any improvement over present methods is beneficial. In addition to high costs of training, consideration must also be given as to the nature of damage that could be caused (in loss of materials, time, etc.) for those people accepted who will be job failures (Anastasi, 1976).

Biographical Data as Predictors

Additional interest in the present study is the inclusion of as much information as is available on

pilot applicants. The potential exists for biodata (or interchangeably - background data), collected by way of application forms to predict outcome in pilot training above that determined from present methods (i.e. ability/aptitude) alone.

The USAF routinely collects background data on all applicants. In addition to test scores, the following types of data are collected: date and place of birth, gender, race, ethnic group, religious denomination, legal residence, academic education level, grade point average, marital status, number, age and sex of dependents, university attended, college major, foreign language proficiency, place from which ordered to active duty, citizenship status, mental category, physical profile, special qualifications (calculus or computer academic hours), personal interview data which includes applicant communication skills, and officer career field preference.

Any incremental validity resulting from the

addition of background data can be subjected to a utility analysis (Cascio, 1982; Schmidt, Hunter, & Dunn, 1987). The gain (or loss) of utility beyond that found with the Pilot Composite alone can be determined.

Previous research into background items has shown that the experiences and events that help shape the development of an individual can successfully predict outcome in a variety of situations. For example, Hunter and Hunter (1984) conducted a meta- analysis of 33 studies predicting training success using some commonly used alternatives to ability tests. Most of the studies involved using the criterion of success or failure in training. The average validity between training success and biodata was 0.30.

Background data generally do not rely on attitudes and feelings as much as on historical events in the person's past. This results in more honest answers (obtained from application blanks) particularly for verifiable events (Owens, 1983).

Under this assumption, information obtained from historical personnel records should predict as well as information derived from a specific biodata response inventory.

Numerous studies have emerged over the years concerning the use of background items as predictors of success in various training and employment situations. College success (Anastasi, Meade & Schneiders, 1960), engineering (Kulberg & Owens, 1960), career aspiration of physical scientists (Albright & Glennon, 1961), management talent (Laurent, 1962), success of creative research personnnel (Buel, 1965) and executives (Rawls & Rawls, 1968), as well as Navy divers (Helmreich, Bakeman, & Radloff, 1973), and Air Force pilot candidates (Roomsburg, 1988) have been areas investigated. All of the studies demonstrate that biodata are valuable as a source of predicting future behavior based on previous behavior. A renewal of interest in background information

(Mumford & Owens, 1983) in the selection literature creates a viable research area to couple with utility analysis.

An additional research opportunity exists in the present study to investigate whether the validity of selected background items is stable over time. The evidence provided for stability of background variables by Brown, in 1978, was disputed by Hunter and Hunter (1984) who contributed his finding to large sample size (N=14,000) which would produce high statistical significance for even small correlation. Hunter and Hunter state that rather than looking at statistical significance (as Brown did), the proof of stability over time involves looking at the actual validity. The present study involves cross validation procedures of background item validity coefficents over time. The results may provide evidence for the stability of background items.

The Selection Environment

Some comments concerning the uniqueness of the military aviation selection environment need to be made. Ironically, applicants are not selected for pilot training, per se. Applicants who eventually enter UPT are selected through procedures which qualify them for admission to one of the three USAF commissioning programs. Each of the three commissioning programs have different selection procedures and pilot goals.

Commissioning Programs

The United States Air Force Academy (USAFA) is a four-year baccalaureate degree awarding institution. Upon graduation, the students are commissioned as second lieutenants. USAFA students are not required to take the AFOQT as a qualifier for commissioning due to the military educational environment of the program.

The USAFA has a light plane training program not unlike the OTS FSP described earlier. Selection

is extremely competitive and commissions are into the Regular Component of the USAF. The USAFA service commitment is a minimum of five years post-graduation active duty and longer if UPT is completed. The USAFA represents the most inflexible method of commissioning due to the length and relatively stable production of a given number of pilot candidates per year.

The Reserve Officer Training Corps (ROTC) selects individuals who then attend civilian degree granting institutions. Upon graduation from these colleges, the individual is commissioned a second lieutenant in the USAF Reserve forces with many being placed on extended active duty for at least four years or longer if pilot training is completed. The individuals who take the AFOQT prior to entry into an ROTC program are made up of people who are either applying for a four-year scholarship (i.e. high school graduates) or one of the shorter scholarships (two or one year) which indicates at least some college for the applicant. AFOQT scores

are sometimes lower than those found with OTS applicants since presumably ROTC testing can occur up to five years before entry into UPT. Therefore, ROTC testing applicants often do not have the advantage of completed education realized by the majority of OTS applicants.

ROTC pilot cadets attend either Pre-Flight Instruction Program (PIP) or the Flight Screening Program (FSP) prior to attending UPT. PIP is a contracted civilian flight instruction program given at an airport near the ROTC detachment. ROTC is a more flexible commissioning method than the USAFA due to the differences in lengths of awarded scholarships.

The final discussion of the commissioning methods concerns OTS, the most flexible commissioning program, briefly described earlier. Great numbers of pilot candidates can be produced through OTS in a very short time (90 days). Likewise, fewer numbers can be accepted into OTS if

the need for pilots decreases. The subjects of the present study were restricted to OTS graduates who were on initial active duty at the time of entrance into UPT. This restriction limits the sample from the overall OTS graduates for the years under investigation. The study sample is assumed to be a random sample of all the OTS graduates who attended UPT during the years of 1984 through and including 1987. The overall number of OTS-UPT trainees for the relevant years include officers who were commissioned through OTS but due to intervening events were delayed in actually entering UPT. Some of these events, were cross-training from other rated career fields such as helicopter pilot, or fixed-wing navigator, and other non-rated support fields. More confidence in the underlying assumptions involved in calculations of selection ratios and base rates results from limiting the present sample to only OTS commissioned candidates on initial active duty.

Even limitation of subjects as just described

does not insure precise determination of estimates. For instance, one would have to individually track a great number applicants to arrive at a precise selection ratio for any given year. This is due to the fact of the variable lag time between pilot selection hurdles. Since this was unfeasible for the present (and most likely any) study the best estimation is the use of the numbers available from the OTS Recruiting Service. The time between hurdles for OTS pilot candidates is generally very short or assumed to be at least within the same year.

Compensation and Comparable Worth

All USAF officers receive pay based upon rank not on type of job performed. Although, aviation incentive pay (flight pay) is paid to pilots, the difference between pilots and non-pilots of the same grade is $125.00 per month. This additional amount is not indicative of the difference in the amount invested for pilots or their "worth" when fully

trained.

The worth of an employee can be interpreted as the contribution made by the individual to the organization. Cascio (1982) states that the contribution of the individual to an organization is not the amount of investment in that individual that the organization has made. The contribution made by the employee to the organization can be estimated from the increase in overall mean scores on the predictor used to select the individual. The higher the mean score - the higher the level of ability and presumably the higher the contribution (and generally profit-making) to the organization.

This definition of the usefulness of the test (i.e. the ability to "measure" the potential contribution of the individual avails itself to employment situations where productivity is easily definable and results in an objective criterion measure. In the situation of military pilot selection the level of productivity is not as obvious and regardless of its value does not result

in clear cut profit.

Utility analysis takes on a different meaning when applied to USAF pilot selection. In this situation, the training situation is the best criterion to establish the utility of a selection measure since the ultimate criterion, discussed earlier is so indefinable. This necessarily translates the criterion of higher profit into investment costs. Although some pilots (such as transport and weather, etc.) utilize their training on a daily basis, combat pilots concentrate on remaining proficient in preparation for demand of their specific skills during wartime. It is conceivable that a fully qualified fighter pilot could represent an investment of one million dollars and never be utilized in that capacity during his/her military career. This situation could be construed negatively and criticized as wasted investment costs. Regardless of the costs, current national policy dictates that the alternative to

maintaining a proficient combat force is more defensible than any possible complaints concerning loss of investment.

The situation just described causes problems in utility analysis when applied to the military pilot selection system. When SDy is estimated tc be 40 percent of mean wages, the difference between a pilot and any other non-pilot officer, would be $1500.00 per year. This obviously is not indicative of the "worth" of a fully qualified pilot to the USAF.

Due to hidden costs, conservative estimates are desirable in utility analysis (Schmidt, Hunter, McKenzie, & Muldrow, 1979). However, low estimations resulting from failure to account for the uniqueness of a particular employment situation is arguably inappropriate also.

Different measures could be employed to attempt to differentiate the "value" of different jobs, including that of a pilot, in the USAF. A "criticality" factor, undoubtedly subjective, which

would denote individual responsibility for the various weapons systems could be used. A more tangible alternative, although not perfect, would be to determine training costs of different career fields.

Since no other officer training exceeds that for a pilot, it would be logical to assume that within- organization training programs could be measured on a point scale reflecting the amount of investment for particular training programs. The amount of USAF monetary investment for different programs would be one method to assign values on a point scale. Direct duty assignment officers (no training school) could be assigned a rating of 1 and officers completing pilot training assigned the value of 10. Other training programs would fall between the 1 and 10 ratings depending of the cost of the particular training program.

As stated before, the focus of the present study is on pilot selection not on the USAF

selection program in general so no attempt will be made here to precisely estimate the point estimation for other USAF training programs. Incorporation of a term, say TF, representing a "training factor", to account for the differences in investment based on training would result in a more accurate within-organization utility estimate.

Accordingly, the utility for the purposes of the present study is derived from the following formula:

$$U/selectee = TF(r_{xy1} - r_{xy2})\ SDy\ \phi/p - (C_1 - C_2\ /p)$$

Where:
- U = gain in utility from use of the selection device
- TF = training factor
- r_{xy1} = validity coefficient between predictor-criterion
- *r_{xy2} = validity coefficient between predictor-criterion of comparison selection method
- SDy = standard deviation in performance in dollars
- ϕ = the ordinate height of the normal curve at p
- p = ratio of selectees to applicants
- C_1 = cost of testing one applicant with device
- *C_2 = cost of testing one applicant with device

*When determining the utility of only one selection method, r_{xy2} and C_2 have values of 0.

The following hypotheses are made based upon the research findings reviewed above:

H1. The use of AFOQT Pilot Composite scores in the prediction of UPT outcome results in a substantial dollar criterion utility.

H2. Background data can be used to successfully discriminate between those OTS UPT candidates who eliminate and those who graduate.

H3. The use of background data in the prediction of UPT outcome results in a significant increase in utility over that realized from the use of the AFOQT Pilot Composite scores alone.

H4. Utility analysis formulas provide a better representation of the usefulness of the AFOQT than the Direct Method currently used by the USAF.

CHAPTER II

METHOD

The primary focus of the present study is the utility of the AFOQT as a function of variable selection ratios and base rates across years. A secondary focus is the increase in utility of the AFOQT after the addition of background data. The following method section will be divided into sections (Phase I and Phase II) to discuss the two areas of interest. An additional section (Phase III) will describe other relevant sub-analyses.

PHASE I

Criteria:

The criteria measures were UPT outcome (coded 1 for graduation and 0 for elimination each year (1984 through 1987).

Predictor Measures:

Predictor measures were AFOQT Pilot Composite

scores for each year (1984 through 1987) under investigation.

Subjects

A total of 1550 OTS graduates on initial active duty were selected from the USAF HRL database. All subjects had taken the AFOQT prior to their selection for OTS. All subjects were graduates of a four-year approved college program and all held grade of second lieutenant. Some subjects had completed graduate school before attending OTS and UPT.

Procedure

#1 Discriminant Analysis (#1DA):

The total sample (N=1550) was divided into four year groups based upon the date of entry in UPT (1984 N=408, 1985 N=381, 1986 N=410, and 1987 N=351).

For each of the year groups discriminant analysis was used to derive discriminant function score (DFS) correlation coefficients between AFOQT

Pilot Composite scores and the UPT performance outcome criterion. The coefficients derived were then used in the following analysis.

#1 Utility Analysis (#1UA):

A modified Schmidt, Hunter, McKenzie, and Muldrow (1979) utility equation was used to determine the total utility of the AFOQT as a selection device. Correlation coefficients derived in #1DA as well as other calculated values described below were used. This analysis addressed the hypothesis that use of the AFOQT results in a significant increase in productivity as measured by the number of successful pilots, as well as a substantial increase in dollar utility using this selection method over random selection.

Additional Equation Terms

1. A PRIORI APPLICANT POOL: The original number (per year) of OTS-UPT applicants.

2. SELECTION RATIO: The ratio of OTS-UPT

selectees to the original number of applicants (per year) to OTS-UPT are given in the following table:

Year	Applicants	Selected	Selection Ratio
1984	3420	996	.2912
1985	2652	1185	.4468
1986	1578	812	.5146
1987	1890	764	.4042

Table 1. A Priori Population Selection Ratios*

*(Obtained from HQ USAF Recruiting Service)

3. SDy (Standard Deviation of Performance in Dollars): Calculations of 40 percent of the annual wages of a second lieutenant under two years of service are given in Table 2. The wages included, base pay, basic allowance for subsistence, basic allowance for quarters (single rate) and aviation incentive pay.

Year	Base	Flight	BAQ	BAS	SDy
1984	1143*	125*	232.50*	102.10*	7692.48
1985	1188	125	238.50	106.18	7956.86
1986	1224	125	245.70	109.37	8179.54
1987	1260	125	253.20	112.65	8404.08

Table 2. Calculations of SDy

*All dollar amounts were derived from the Air Force Almanac published by the Air Force Association for the year indicated

4. COST: $12.12 per applicant tested. This figure was derived from the Bortner and Ree work (1977). Their comprehensive research involved the cost estimations of all relevant research, development, support and operational components of the AFOQT. This figure applies to each applicant tested whether tested individually or in groups since the figure was derived by dividing the annual total of the AFOQT cost components into the total annual number of applicants.

The base rate was determined to be 0.7424.

This value is the average of the proportion of pilots considered successful under previous forms of the AFOQT (years 1970 through 1980) to the total number of OTS-UPT entrants for the respective years. Although the versions of the test are correlated, this value represents the population base rate while the base rate derived from use of the current version is designated as the present employee base rate.

PHASE II

Criterion:

The criterion measure was overall UPT outcome (coded 1 for graduation and 0 for elimination) for the entire sample.

Predictor Measures:

Predictor measures were biodata items that were collected during the selection process. Items that were included were flying hours obtained before OTS, gender, possession of a private pilot's license

(PPL), ethnic group, race, marital status, geographical area from which the individual entered the USAF, age upon entry into UPT, type college attended, level of education, academic specialty, height, weight, number of dependents, religion, number of computer and calculus hours taken, declared preference for pilot training, and citizenship status.

Subjects

All of the 1550 subjects from Phase I were utilized for Phase II. The majority of the subjects had provided biodata information during the selection process. All subjects were graduates of a four-year approved college program and held the grade of second lieutenant. Some had completed graduate school prior to entry into OTS and UPT.

Procedure

#2 Discriminant Analysis (#2DA):

The total sample (N=1550) was divided into

development (N=789) and validation (N=761) groups based on year of entry into UPT. The development group consisted of all entrants into UPT for years 1984 and 1985. The 1986 and 1987 groups were combined for the validation group. The division method allowed for investigation of stability over time as well as the other planned analyses.

For the development-group and validation-group, separate discriminant analysis was used to derive discriminant function score (DFS) correlation coefficients between biodata responses and the UPT performance outcome criterion. These correlations were between biodata items retained in a stepwise reduction method and the UPT outcome. These items were then applied to the validation-group and correlations were computed between the validation-group and the development-group derived items.

Conversely, correlations were derived between biodata items retained in the validation group and UPT outcome. These items were then cross-applied and correlations between the validation-group

derived items and UPT outcome using the development-group were obtained. The average of the two cross- validated correlation coefficients was then computed. The whole sample was then used to assess the background data items and the resultant correlation coefficient was used in further analyses. This analysis addressed the hypothesis that biodata can be used to differentiate between those people who succeed in UPT and those who fail.

#2 Utility Analysis (#2UA):

A modified Schmidt, Hunter, McKenzie, and Muldrow (1979) utility equation was used to determine the incremental utility of the addition of biodata. The correlation coefficient derived in #2DA as well as other calculated values described below were used. This analysis addressed the hypothesis that use of biodata results in a substantial increase in dollar utility using this selection method in addition to the AFOQT.

Additional Equation Terms

1. A PRIORI APPLICANT POOL: Calculated as in Phase I.

2. SELECTION RATIO: Calculated as in Phase I.

3. SDy (Standard Deviation of Performance in Dollars): Calculated as in Phase I.

4. COST: $10.00 per applicant tested. This figure is based on references in the literature, concerning paper-and-pencil tests (Schmidt, Hunter, McKenzie, & Muldrow, 1979).

5. BASE RATE: 0.720 - Calculated from the total sample (N=1550) which demonstrates the relationship of successful trainees (N=1116) to the total number of OTS-UPT trainees in the sample.

PHASE III

Sub-Analysis #1:

Classification decision tables were generated to compare the number of correct and incorrect selection decisions made with the use of the AFOQT Pilot Composite score and background data as

predictors for the total sample (N=1550). From these results, a method will be demonstrated to estimate the full validity model (as represented in Figure 1) from classification information, base rates/<u>/\</u>base rates and selection ratios/<u>/\</u>hiring rates determined in the present study.

<u>Sub-Analysis #2</u>:

Pilot Composite scores were transformed into equal-N decile groups for the entire sample (N=1550). Base rates for each decile were calculated and a Direct Method cost-benefit analysis was conducted. The Direct Method results were then compared to a total sample utility analysis based on the modified Schmidt and Hunter formulas.

CHAPTER III

RESULTS

The results of the data analyses will be presented in separate sections corresponding to the analyses Phases presented in the previous chapter. Phase I addresses Hypothesis #1 analysis results, Phase II presents Hypotheses #2 and #3 results, while Phase III will describe the results for Hypothesis #4.

PHASE I

The total sample was divided into four separate groups based upon UPT class entry dates. Separate utility analyses for each year group were conducted to demonstrate interactions between the various equation parameters which differed as a function of year.

A discriminant analysis procedure was conducted on each of the groups to derive discriminant function scores (DFS) from the Pilot Composite scores. This discriminant analysis procedure

resulted in a correlation coefficient between UPT outcome (coded 0 and 1) and the discriminant scores for each year. The predicted group classifications for each year were also based on these DFS. These results are given in Table 3.

Since discriminant analysis is comparable to regression analysis, cross-validation in the univariate predictor case is unwarranted. Attempts to apply the weights derived in a development group to a validation group will result in only a linear transformation of the validation group's original correlation and vice versa. Therefore, in Phase I, cross-validation was not performed and each year group was entered into the analysis as a whole. The size of the low correlations found were attributed to the acknowledged restriction in range and relatively large sample size.

Even though the AFOQT Pilot Composite is the major selection device for UPT, other factors can contribute to the selection decision since final decisions are made by selection boards. In

addition, even if correlations were completely corrected for restriction of range, some error in prediction would exist.

If all other factors and the error were quantified and used for classification, the result would be to classify all in the pass group. This is evident when one considers all 1550 people were predicted to pass by the selection board. It is understood when making this statement that the board members are probably aware that some trainees will fail but the pilot goals which establish the number selected take this into account not the board. It is probably unnecessary to say that the board selects based on who they believe should pass.

Individual case prediction is based on the DFS derived from the Pilot Composite scores alone. These predictions are then compared to actual UPT outcome and cases are classified based on how likely the case is to be assigned to either the pass or fail group.

If the correct classification (high hits) can be improved over actual group assignment, then the prediction improves on the base rate. Rather than considering the actual pass rates in the sample as representative estimates of the population base rates, they can be considered as the high hit rate of the selection board.

It might be noted that the overall correct classification rate is not the definition of /\BR. Technically, definition of /\BR in these terms is high misses plus high hits divided by the total number selected. This fact sets /\BR apart from the traditionally defined success ratio. The success ratio is defined as the high hits divided by the total number selected.

By using the Pilot Composite alone, one of the selection board ($/\BR_{SB}$) factors has been partialed out into a separate selection hurdle. $/\BR_{PC}$ (Pilot Composite) can now be defined as as the high hit rate given in Table 3. When the Pilot Composite score was used alone it predicted as well as the

total board in three of the four year groups. It is possible that for those years, the board relied more heavily on the Pilot Composite since it appears that the Pilot Composite alone predicted as well as the board.

This provides evidence for the explanation that the board, even when considering other factors, predicts no better than the Pilot Composite alone. An alternative explanation could result from differences in the applicants. The 1984 year group had the lowest overall Pilot Composite mean score (71.81) which could mean that the board considered other factors equally with the Pilot Composite. The lower scores should have resulted in a larger range which would account for the greater correlation found in the 1984 year group between the Pilot Composite and UPT outcome. This prediction did not continue in the final base rate results for the year group. If the result is that the board considered other factors more heavily for that year, the Pilot

Composite may have been less predictive in that instance but normally predicts at or above the total board decision. Results for #1DA are given in Table 3. The actual present sample UPT base rates for each year are designated $/\backslash BR_{SB}$.

Year	r	$/\backslash BR_{SB}$	$/\backslash BR_{PC}$
1984	.2193	77.70%	75.98%
1985	.1123	81.10%	81.10%
1986	.1265	70.73%	70.73%
1987	.1572	56.98%	56.98%

Table 3. Results of Discriminant Analysis #1.

The present model utility formula presented in Chapter I was used to test Hypothesis #1. The DFS correlations of the Pilot Composite with UPT outcome as well as other parameters previously defined were combined to compute the utility per selectee for each year group given in Table 4.

YEAR	Utility Per Selectee
1984	U = 10 (.2193) (7692.48) (.3423/.2912) - (12.12/.2912) = $19788.28
1985	U = 10 (.1123) (7956.86) (.3951/.4468) - (12.12/.4468) = $ 7874.48
1986	U = 10 (.1265) (8179.54) (.3987/.5146) - (12.12/.5146) = $ 7993.15
1987	U = 10 (.1572) (8404.08) (.3876/.4042) - (12.12/.4042) = $12638.66

Table 4. Results of Utility Analysis #1 (#1UA)

Variations between the equation parameters can be seen in the individual formulas in Table 4. The highest correlation coupled with the lowest SDy and selection ratio values (1984) results in the comparatively highest utility per selectee. The lowest utility results when the lowest correlation is coupled with comparably high selection ratio (1985). More obvious is that the level of r is

comparable to the level of utility (i.e. highest $\underline{r}$ equals the highest utility and the lowest $\underline{r}$ results in the lowest utility), regardless of other parameter variation in the year groups. Beyond this, no further patterns are discernable.

Total utility for a test for any year can be defined as the utility per selectee times N, where N is the number selected per year. The combined and averaged utility estimate of the Pilot Composite, for the present sample, resulted in a substantial dollar criterion utility amount of approximately 18.7 million dollars. This finding substantiates Hypothesis #1. All Phase I results are provided in Table 6:

UPT YEAR	$/\BR_{SB}$	SR	r	UTIL/SEL	$/\BR_{PC}$
1984	77.70%	.2912	.2193	19788.28	75.98%
1985	81.10%	.4468	.1123	7874.48	81.10%
1986	70.73%	.5146	.1265	7993.15	70.73%
1987	56.98%	.4042	.1572	12638.66	56.98%

Table 6. Phase I Results

PHASE II

Hypotheses #2 and #3 concerning the addition of background data into the selection process for UPT were addressed in this Phase. The determination of any increase in validity above that found with the Pilot Composite alone and analysis of any gain in utility, with the addition of background data to the selection process, was of interest.

The total sample was divided into two groups based upon entry year into UPT. The first group (development) consisted of those subjects entering UPT in 1984 and 1985 (N=789) while the second group (validation) consisted of those entering in 1986 and

1987 (N=761).

This method achieved the results of meeting the independence of sampling time requirement by Siegel and Lane (1983) who state that for true cross-validation, subgroups of the sample must be collected at different times. Herriott (1988) also employed this method of cross-validation to achieve independence of samples. In addition, the results from this method allow the investigation of background item stability over time through cross-validation of the biodata items.

Discriminant analysis was used to differentiate (based on a combination of background data items) between those who succeeded in UPT and those who failed. Investigation of background data was conducted in an attempt to counter the restriction in range problem of selectees. Utilizing predictors with a more variable range of responses (due to the fact that the subjects were not systematically selected on the background items) would be one

method of countering the restriction in range. Moreover, regardless of possible restriction in range based upon nonsystematic selection on the background data, this method could provide a means to quantify the information by a more systematic means.

A stepwise variable reduction method based upon Wilks Lambda was applied to the development group to retain the most accurate combination from an original set of 41 background data predictors. This resulted in a statistically significant correlation of 0.3213 ($p < 0.0000$) between seventeen retained items and UPT outcome. The 41 original items were then used in a Wilks Lambda stepwise variable reduction method in the validation group and resulted in a statistically significant correlation of 0.2232 ($p < 0.0071$) with 14 items retained.

The seventeen items retained from the development-group were then applied to the validation-group. A smaller and not statistically significant correlation of 0.1723 ($p < 0.2123$) was

obtained in this cross-validation. The opposite result occurred when the 14 retained items from the validation-group were applied to the development-group. An even higher cross-validated correlation of (0.2253, $p < 0.0002$) was obtained than was originally derived in the validation-group. This could be due to the larger N found in the development-group or as a function of the lower mean obtained by the those who failed UPT in the development-group.

Even though the development to validation cross items did not reach statistical significance, usable items resulted for classification purposes. The classification resulting from even the low $\underline{r}$ between the DFS and UPT outcome resulted in better discrimination between the two UPT groups (pass/fail), than did the DFS based on the Pilot Composite.

In addition, seven of the 17 items retained in the development group were the same as seven of the

14 items retained in the validation group. This was interpreted to mean that these seven were important items to both samples even though in the development to validation cross, the combination with other items did not reach statistical significance.

In an attempt to demonstrate this further, the seven were forced back into each group to utilize a larger portion of the samples. Since data were not available for all of the subjects, a reduced number in the development group (N=523) and the validation group (N=601) were used to compute the correlations and retain items from the original 41 item set. (Cases are excluded when they are missing one of the discriminating variables. The higher the number of original variables, the more likely cases will be excluded).

The result of the forced entry of the seven variables was a development group (N=784) correlation of 0.2124 (still statistically significant: $p < 0.0000$) and an increased validation group (N=758) correlation of 0.1514 which

was significant ($p < 0.0147$). Since the larger correlation in the development group could have resulted from the larger N, the two correlations were then transformed to Fisher Z values and averaged. The Z to r transformation resulted in a cross-validated r of 0.180 ($p < 0.001$).

Double cross-validation of the correlations was not warranted as the seven variables are considered as a group in one linear combination. This results in the univariate predictor situation discussed earlier. The statistically significant correlations found in the subsample averaging of correlations provide evidence that biodata can be used to differentiate between those who succeed in UPT and those who fail as well as suggesting that the particular items are stable over time.

Following this reasoning, the seven items were applied to the whole sample (N=1550) for use in further analyses. The total sample resulted in an incremental validity coefficient of 0.1798. Though

lower than either the development or validation group, due to size of the sample, very high statistical significance was achieved ($p < 0.0000$). An attempt to further reduce the seven items in the total sample by a Wilks Lambda reduction method, as expected, was unsuccessful This retention of the seven items in the total group demonstrated further evidence that the items, in combination, were important predictors. The high hit rate resulting from this total sample analysis was defined as $/\backslash BR_{BIO}$.

The seven retained items addressed various background factors. People who majored in engineering were more likely to succeed than those who declared no specific major. Candidates who were from the far southwestern states (California, Nevada, New Mexico, Arizona, Colorado and Utah) were more likely as a group to succeed in UPT. Flying experience before UPT was a positive influence on success in UPT. Those people who declared no specific ethnic background were more likely to

graduate as were those who were married. Overall, the younger cohort of trainees was associated with success in UPT.

The findings of this analysis suggest that a biodata combination can be used to differentiate successes and failure in UPT. The results also provide evidence that these items (in combination) are stable over time.

Cross-validation is generally thought of in the context of cross applying weights to variables in a validation group from the development group. Campbell (1983) discusses cross-validation as a method to test decision rules by validating the rules on a group not used to develop the rules. It is in this context that the emphasis was on cross-validation of biodata items. The main interest was in whether the items would significantly validate (be retained) and whether they were stable over time. Both were shown to occur.

As stated earlier, the total sample was

combined and a discriminant analysis using those items which were retained for both the development and validation groups were used in the final analyses of Phase II. It is realized that use of all the subjects generally results in an upward bias. On the other hand, the correlations resulting from the original two cross- validations are biased downward because they are computed from less than the total number of subjects (Campbell, 1983). The possible upward bias was believed to be more acceptable, if present, due to the restriction in range of the current sample. The upward bias would be countered by this restricted range.

Although the restriction in range was on the Pilot Composite scores and presumably not the background data, background data significance could in fact be related to the Pilot Composite scores. If so, the bias would be downward even for the larger sample size.

The overall gain in utility with the addition of a background data hurdle (ΔBR_{BIO}) was 21.46

million dollars. This represented a translation of the incremental validity, found with the addition of background data, into a utility gain of 2.76 million dollars above that of the Pilot Composite alone. Stated in other terms, this analysis demonstrated an $1800.00/per selectee gain over the utility from use of the Pilot Composite alone.

An analysis to demonstrate the effects of adding the background data at the same hurdle as the Pilot Composite was also performed. The Pilot Composite score was entered into a discriminant analysis with the seven retained background variables and applied to the whole group (N=1550). This analysis resulted in a relatively large correlation of 0.2256 ($p < 0.0000$) which exceeded both the Pilot Composite alone and the background data alone. The high hit results from this analysis produced the measure of $/\backslash BR_{PCB}$ (Pilot Composite plus background data).

If background items would have been

incorporated at the same hurdle ($/\backslash BR_{PCB}$) and thus considered with the Pilot Composite simultaneously and systematically (rather than background data considered alone as an incremental validity gain), the overall utility would have reached 26.43 million dollars (over random selection).

PHASE III

Sub-Analysis #1

The following classification results were obtained from an all subjects classification based on the Pilot Composite alone, biodata alone, and both combined:

	PILOT COMPOSITE		BIODATA		COMBINED	
		Predicted		Predicted		Predicted
Actual						
		Fail/Pass		Fail/Pass		Fail/Pass
Fail	434	6 428		9 425	24	410
Pass	1116	1 1115		7 1109	15	1101
hit rate		72.32%		72.13%		72.90%

$/\BR_{PC}$ (71.94%) $/\BR_{BIO}$ (71.55%) $/\BR_{PCB}$(71.03%)

Table 7. Phase III Classification Comparison

The background data alone actually resulted in a less accurate overall hit rate than the Pilot Composite alone, but the discrimination between cases is more accurate in the direction of interest (more of those who were predicted to fail actually failed). The discrimination is even more apparent when Pilot Composite and biodata were combined. The overall hit rate with the combination of the background and Pilot Composite at the same hurdle is also greater than either the background or Pilot Composite alone. The results of the $/\BR$

calculations appear to contradict the concept involved with Delta Base Rates as developed in Chapter I. However, a graphic demonstration at the end of the chapter will show how the Delta Base Rate actually increases as an inverse function of hiring rate.

Sub-Analysis #2

The Direct Method of cost accounting was compared with the Schmidt-Hunter adapted formula. The total sample was treated as an "applicant population" for the comparison by rank ordering the subjects on Pilot Composite scores. The total sample was then segmented into decile groups of approximately 10 percent each to aid in the comparison.

For this analysis overall averages of the total "population" for formula parameters were used. A correlation of 0.1562 (based on actual correlation in the "population") and SDy of $8058.24 (based on 40 percent of average pay over these four years)

were defined as the population parameters.

The figures in Table 9 were used for comparison between the two methods. Table 9 adds the individual deciles Ns to the left for ease in computation and comparison. The CS line represents the cutting score defined for the present comparison.

	1	2	3	4	5	6	7	8	9	10
Fail	434	363	312	269	223	181	138	96	58	23
Pass	1116	1026	915	815	704	593	469	349	251	114
Total	1550	1389	1227	1084	927	774	607	445	309	137
△BR	.72	.74	.75	.75	.76	.77	.77	.78	.81	.83
						CS				

Table 9. Top Down Selection Data (SR = 0.40).

The terms of the comparison were defined as the utility gain measured by both methods when the cutting score was set at a score which resulted in the top 40 percent being selected (SR = 0.40).

The Direct Method estimates the total cost of testing to be $18786.00 (1550 X 12.12). Of this total amount, testing funds for 943 rejected candidates (1550-607) resulted in a loss of $11429.16. 296 persons (434 minus 138) were identified as correctly rejected eliminees with the 60 percent cutoff. This resulted in a savings of $19.83 million (296 X 67000.00). Actual eliminees of those selected resulted in a loss of $9.3 million (138 X 67000). The total savings based on the Direct Method of cost effectiveness results in $10.58 million. The actual Direct Method cost-estimation formula would be:

DMCE:

(296)(67000) + (138)(67000) - (943)(12.12) = $10.58

Gain in utility based on the Schmidt-Hunter adapted formula is computed by:

S&H(A):

10 (.1562) (8058.24) (.3863/.4000) - (12.12/.4000) = 12125.57/selectee OR 12125.57 X 607 = $7.36 million

A size-only comparison demonstrates that the Schmidt-Hunter adapted formula value for utility gain with the Pilot Composite, though substantial over random selection, is approximately 3.2 million dollars less than that estimated by the Direct Method of cost estimation. Hypothesis #4, however, is not stated in terms of size but in terms of whether or not the Schmidt-Hunter adapted formula provides a better representation of the usefulness of the Pilot Composite.

Both formulas consider test scores and selection ratios. The Direct Method however, is primarily concerned with loss of investment and does not consider the value of successful selectees. It involves the estimation of savings based on valid rejects and false rejects as well as the loss in rejection test funds. The Direct Method implicitly

accounts for investment costs and explicity disregards specific values of correlation between the predictor and UPT outcome while considering base rates. There is no evaluation of the costs associated with false rejects and their potential value had the predictor been able to provide better classification.

On the other hand the Schmidt-Hunter adapted formula relies on a combination of the correlation coefficient, the SDy, the test score (represented by ϕ/p), the selection ratio, and testing costs but not base rates. The Schmidt-Hunter adapted formula does not address loss of investment, per se, but considers the value of successful or "productive" employee. The mean output (represented here by 40 percent of pay) increases as a function of the correlation and selection ratio. The higher the correlation and the lower the selection ratio, the higher the gain in utility.

Since Table 10 also represents a transformation

of the sample data into a rectangular distribution of equal-N deciles the utility found for total group selection (SR = 1.00) is negative and reflects only testing cost loss. There is no gain in utility found when all applicants are selected as reflected in this representation of a "population".

In this comparison between the two methods of analysis the Schmidt-Hunter adapted formula results in positive utilities for all selection ratios other than 0. The Direct Method will result in negative values for any selection ratio of more than 50 percent. In fact, based on the Direct Method of cost estimation, selecting no applicants will result in a gain of 29.06 million. The Schmidt-Hunter adapted formula results in a fairly normal distribution of utility values with a median value of $7.75 million at SR = 0.50.

% Selected	N	DMCE	S&H(A)	/\BR
0	0	29.06	0	0
10	137	25.98	3.01	.83
20	309	21.29	5.43	.81
30	445	16.21	6.47	.78
40	607	10.58	7.36	.77
50	774	4.81	7.77	.77
60	927	- .82	7.49	.76
70	1084	- 6.93	6.76	.75
80	1227	-12.74	5.39	.75
90	1389	-19.52	3.39	.74
100	1550	-29.08	- .02	.72

Table 10. Comparison Of Utility Estimation Methods (In Millions) In Top-Down Selection.

The information in Table 11 is provided for use in a top down selection process based on actual sample decile mean Pilot Composite Scores. This represents a rank ordered top-down selection process using the present sample.

ΔBR_{PC}	Pilot Composite Score	SR	Decile
.83	97	.10	10
.81	93	.20	9
.78	87	.30	8
.77	84	.40	7
.77	79	.50	6
.76	74	.60	5
.75	69	.70	4
.75	63	.80	3
.74	55	.90	2
.72	39	1.00	1

Table 11. Top Down Selection Score Cutting Scores and ΔBR_{PC}

For instance, if only those people who scored above the 87th percentile had been selected (SR of 0.30), 78 percent would have been predicted to succeed. Improvement over the population base rate would have been achieved by allowing the top 80 percent to enter UPT.

If all persons in this "population" would have been allowed to enter UPT, 28 percent of them would have failed. If only 10 percent would have entered there would have been a reduction in failures to 17

percent. Although the reduction of failures in this situation is significant because of the costs involved, Table 11 provides evidence for the Taylor and Russell (1939) and Brogden (1946) agreement that improvement in the number of successful people hired is difficult when the population base rate is extreme.

The information in Table 11 demonstrates the prediction from the sample to the population. Assuming this sample is representative of the "a priori" population, the values in the third column represent the selection ratio as a parameter in the population. Eighty-one percent would be predicted to succeed from any random sample from the a priori population of which the top 20 percent were selected.

The results in Table 11 can be used to test the addition of predictors other than those used in the original selection of the present subjects. A graphic display of the relationship between Delta Base Rates, selection ratios, and hiring rates can

be demonstrated to test to addition of background data to the "population" as represented by the present sample. In this representation, the selection ratio, although necessary for utility calculation, is joined by the sample statistic hiring rate defined here as those "hired" as pilots (i.e. UPT graduates).

With these considerations in mind, the following predictive validity models (based on Figure 1) and data from Table 11 are presented. The first interpretation demonstrating the sample to population comparison is given as the Restricted Model in Figure 4. To aid in deciphering the subsequent graphs, lists for terms in each will be provided above the graph.

Figure 4 Terms:

SR=Selectees/Applicants	BR=Successful/Entered
HR=Number Hired/Selectees	SR=HR N=1550
VA=Valid Accepts	FA=False Accepts
VR=Valid Rejects	FR=False Rejects
BR=VA+FR/N(HR)	HR=VA+FA/N

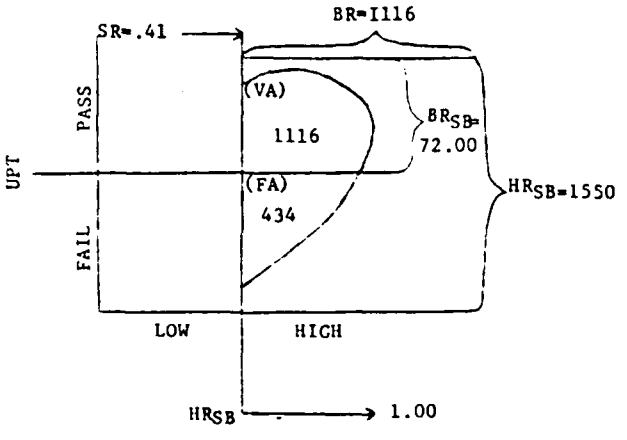

Figure 4. Restricted Validity Model Based on Selection Board Prediction of UPT Outcome

This restricted validity model results from consideration of the selection board decision

applied to the present sample. Notice that Areas 3 and 4 (from Figure 1) are unknown and are set at zero. Even though it is known that 9540 applicants were considered by selection boards from 1984 through 1987, only 1550 subjects in the present sample are from the 3757 that were selected. As such, the true UPT base rate of the non-selectees will never be available and the 2207 other selectees are not presently known for the purposes of this study. Figure 4 thus represents the usual situation in personnel selection. Note that BR and HR are provided both in percentages as well as actual numbers. HR although equal to SR in this situation has the value of 1.00 since all selected at this point of sample to population comparison were "hired".

The graphic representation in Figure 5, made possible with the present sample data, represents full validity model. Since the interest is in comparing the test score data to background data, the Pilot Composite was partialed out of the

selection board base rate. For comparison purposes, it will be assumed that the Pilot Composite has been treated as a separate hurdle beyond the selection board decision. An alternative view would be compare Figure 6 with Figure 4 which would relate the findings of adding the background data hurdle into the selection process to the original group selected by the board with the Pilot Composite retained as a board factor.

Figure 5 Terms:

SR=Selectees/Applicants
HR=Number Hired/Selectees
/\HR=/\FA+/\VA/N
VA=Valid Accepts
/\VA=VA fm new predictor
VR=Valid Rejects
/\VR=VR from new predictor
BR=VA+FR/N(HR)
/\BR=/\VA+/\FR

BR=Successful/Entered
SR=HR
N=1550
FA=False Accepts
/\FA=FA fm new predictor
FR=False Rejects
/\FR=FR fm new predictor
/\HR=/\VA+/\FA/N

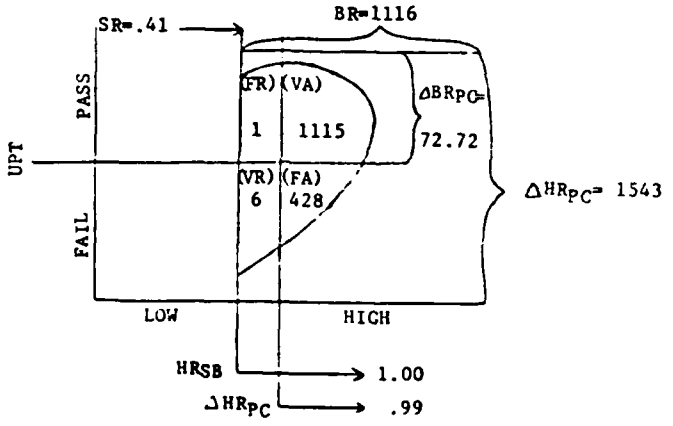

Figure 5. Full Validity Model Based On A Separate Pilot Composite Hurdle

Figure 5 shows that with the addition of another predictive hurdle onto a group of known

employees, the full validity model described by Taylor and Russell (1939) is achieved. The addition of the Pilot Composite results in a $/\backslash BR_{PC}$ of 72.72 compared to the $/\backslash BR_{SB}$ of 72.00, a small but nonetheless important difference. The 72.72 value translates to 1127 "hirees" from the original 1550 selectees (refer to terms for actual computation). Employing the concept of $/\backslash HR$, an alternative comparison would be the entering of only 1543 trainees into UPT to achieve the same 1116 hirees obtained from the present 1550 selectees. Therefore, the use of the Pilot Composite hurdle alone results in either eleven more successful pilots or the same number of original successful pilots but with the savings of associated costs of sixteen less people entered into training.

Figures 6 and 7 demonstrate the background data hurdle and combined Pilot Composite/background data hurdle, respectively. The terms for these last two figures are the same as Figure 5:

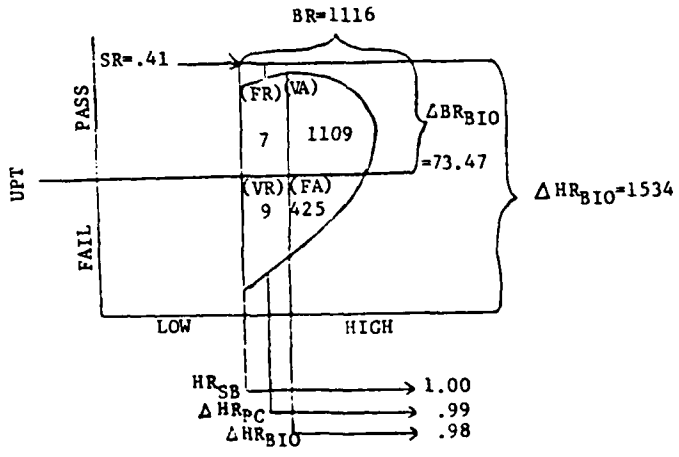

Figure 6. Full Validity Model Based On The Background Data Hurdle

Figure 6 demonstrates that the increased ΔBR_{BIO} of 73.47 and ΔHR_{BIO} of 1534 results in 1139 hirees from the original 1550 selectees or a gain of 12 successful pilots over Pilot Composite alone or 23 over "random selection" with use of background data. Only 1519 have to be entered into UPT to obtain the original 1116 hirees (31 less than random

selection). 15 less people are entered than in the Pilot Composite hurdle considered alone.

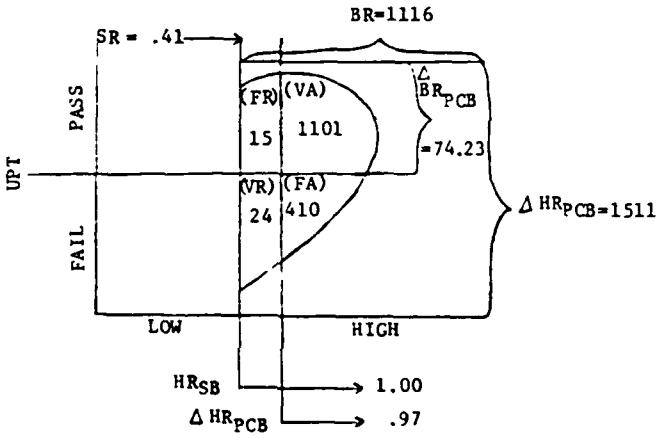

Figure 7. Full Validity Model Of A Combined Pilot Composite And Background Data Hurdle

Finally, Figure 7 represents the most significant results. The use of the Pilot Composite and the background data combined into one hurdle (ΔBR_{PCB}) results in 1151 hirees (an increase of 35) with the original number of entrants or 47 fewer

entrants to produce the original 1116 pilots.

In all of the figures, reference to increase in successful hirees can be interpreted to mean increase in the Delta Base Rate for the given hurdle. This method of deriving the usefulness of a selection instrument discounts the specific correlation between the selection instrument and the criterion however, the effects of the correlation can be seen in the actual numbers within the model areas.

Some contamination from the Pilot Composite to selection board comparison may exist and therefore a caveat is applied in that instance. The background data comparison is assumed to be less correlated with the previous hurdles, at least from a systematic point. The principles underlying the concept and development of Delta Base Rate and the differentiation between selection ratio and hiring rate however, have been demonstrated.

CHAPTER IV

DISCUSSION

The present study demonstrates that the Pilot Composite subtest of the USAF AFOQT is a useful predictor of success in undergraduate pilot training and its use results in substantial dollar criterion utility. The findings also suggest that incorporation of background items, as another hurdle in the selection process, results in better discrimination between pass/fail groups (via a higher correct classification rate and Delta Base Rate) than the use of the Pilot Composite alone.

The inclusion of background items demonstrated incremental validity and increased utility, measured in dollars, above that gained from use of only the Pilot Composite. When combined and measured in terms of utility gain, the Pilot Composite with the systematic quantification of background items exceeded either predictor, alone or in tandem.

The final finding was that the Direct Method of

cost estimation is a useful method to determine the worth of a selection instrument. A limitation to the Direct Method is its retrospective nature. It is generally applied after the fact to judge the value of what the selection instrument was. With no incorporated method to estimate values of savings based on correctly rejected eliminees and losses due to actual eliminees the method is rendered virtually useless in the utility projection environment.

In addition, with the selection ratio selected in the present comparison, the Direct Method resulted in a higher estimate of utility than did the Schmidt- Hunter adapted method However, further calculations showed that had the percentage of subjects selected ranged above fifty percent, only negative utility would have resulted.

Negative utility can be found in the applied selection but generally occurs when the costs involved in the selection process are higher than the value of the identification of selection decision errors. Use of the Direct Method suggests

that concentration on only correctly rejected eliminees results in negative utility above certain levels of the selection ratio even in this situation of high cost decision errors. The ramification of this finding is that negative utility determination often results in the discredit of a selection instrument. A method to consider more of the information available within the selection environment results in perhaps more involved but more informative estimation method.

The Schmidt-Hunter adapted utility estimation method resulted in different estimations of utility from the Direct Method. The greatest gain was achieved at the selection ratio of 0.50. Reduction in utility resulted as the selection ratio became more extreme in either direction. Since the formula is based on parameters that not only can be tested but projected, this method seems to result in the more beneficial system from the standpoint of personnel selection planners. This findings as well

as those previously stated will be discussed more fully in the following sections. The separate hypotheses will form the basis for the discussion.

The test of Hypothesis #1 found that the use of the AFOQT Pilot Composite scores in the prediction of UPT outcome resulted in a substantial dollar criterion utility. The 18.7 million dollar gain in utility from use of the Pilot Composite over random selection was obtained despite the low correlations between the Pilot Composite and UPT outcome.

Historically, the Pilot Composite has resulted in low correlations, so from that aspect the sample is representative of the a priori applicant population. Factors such as restriction in range and self-selection admittedly exist in the present sample as well as the a priori applicant population. These factors result in low levels of correlation regardless of the high levels of significance (which in this case is attributed to the sample size).

Evaluation of the Pilot Composite within the framework of utility analysis was an attempt to

counter low correlations and to provide another measure with with to judge the worth of the Pilot Composite. The results of the present study demonstrate that the Pilot composite remains a useful selection tool in terms of dollar criterion measures. Regardless of the low correlations, the substantial utility resulting from the Pilot Composite warrants its continued use as a predictor of UPT outcome.

A warning is given at this point concerning the actual dollar figures derived from the present research. The training factor (TF) value in the Schmidt-Hunter adapted formula, as defined, results in a utility gain of "18.7 million" dollars. It could be criticized that instead of a 1 to 10 TF level, a 0.01 to 0.10 or 100 to 1000 could have been used which would have resulted in different dollar values of utility. This is true and recognized. Further research would enable development of a more definite training investment point scale to apply to

this type of evaluation and once a standard was set (at least internal for the particular organization) more accurate "true dollar" measures could be made.

However, it is important to point out that regardless of the point scale values, the rank ordering of the comparable worth result still exists. With more detailed information, a more accurate determination of the comparable worth of a pilot compared to other career fields within the USAF can be estimated with this method.

The test of Hypotheses #2 and #3 revealed that use of biographical data demonstrates a discriminative power between groups who succeed in UPT and those who do not. Incremental validity and utility gain from the addition of background data as a separate hurdle in the multiple hurdle selection process was demonstrated. The total amount added to the 18.7 million (resulting from the use of the Pilot Composite score over random selection) was 2.8 million dollars for this sample alone. The total utility for the sample incorporating Pilot Composite

and background data as separate hurdles approached 21.5 million dollars. Even if the exact dollar figures are not considered, the results show that improvement in prediction and utility is made with the inclusion of a systematic background data measurement.

A more important finding was the result of the combination of the Pilot Composite and the background items together. The importance of this finding is not totally in the estimated substantial dollar result. The pilot selection boards determine who enters UPT based on a variety of factors. Although each board has point scales and presumably the candidates are selected from a rank ordered top-down selection procedure, different boards and even different board members may select from a variety of particular selection package data.

The present finding provides a method, and in fact, a group of specific items, that when taken together (1) result in substantial utility and (2)

provide a quantification of what previously may have been a haphazardous collection of data submitted for possible consideration by the board. This is not an inference that the selection has previously resulted from a disregard of systematic selection, but a discernable quantification of factors involved in the selection process would provide an increase of the usefulness the total information available on the candidates.

Further research is warranted to determine the optimum set of background data items. Although the seven retained items demonstrated strong evidence for their use in prediction of UPT outcome, the type of analysis used combined the items into a composite. Together, the items predicted well but item interactions and full interpretation of the underlying meaning of what the composite is tapping requires further research. Additionally, although seven of the present data items have proven valuable, further investigation can be directed at the development of specific background data

collection which would result in even better prediction of UPT performance.

The findings from the test of Hypothesis #4 were mixed, in that depending upon the focus of the investigation, either of the two methods would be useful. In a retrospective analysis and when the selection ratio permits the top fifty percent or less to enter UPT, the Direct Method of cost estimation would result in a defendable respresentation of the usefulness of the Pilot Composite as a predictor.

This is not to say that positive utility is always the ultimate goal in validation of a test. Some tests justifiably result in negative findings and depending on the attribution, may deem the test unuseful. In the present situation, however, the Pilot Composite has historically been proven as a useful predictor. When it is used, overall, less attrition in UPT results. It is counter-intuitive to employ a selection method that results in

negative findings of worth, when measured in dollars, whenever a selection cutoff of fifty percent is reached. If the USAF continually selected a reasonable range less than fifty percent, more credence could be given to the Direct Method.

However, in the present study alone, evidence exists that even the USAF does not have the consistent luxury of extreme selection opportunity. The 1986 year group selection ratio was over this fifty percent point. Had the Direct Method of cost estimation been employed, a possible negative or break-even utility would have likely occurred.

Since true utility can only be estimated, when two methods result in different values it is impossible to prove which is the underestimate and which is the overestimate. The current findings, however, show more support for the Schmidt-Hunter adapted method than the Direct Method for the present situation.

The main concern found in the use of the Schmidt-Hunter formula was the calculation of the

standard deviation of performance in dollars based upon 40 percent of wages. In the situation of the non-profit organization, particularly when the lack of an easily definable "production unit" exists, the 40 percent approximation may not be as useful as in other employment or even training environments. The Training Factor which resulted in the adaptation to the Schmidt-Hunter formulas was meant as a counter to this concern and provided a method for intra-organizational comparisons. The within organizational evaluation of the usefulness of a selection instrument is necessary but the generalizations beyond the organization to one with different characteristics are probably limited. Even considering this, the present study was not an effort to return to the period of validity (or utility) specificity.

The USAF personnel environment is difficult to evaluate in the aspect that wages are based on rank held and not the type of position occupied. In

addition, previous utility analyses have not addressed the issue of estimation when cost of living pay increases are received. Continual utility analysis of an organization which pays these type of raises would result in increased utility results when in reality, the productivity of the individuals may remain relatively constant. To the extent these types of raises do not equal the economic cost of living index, the attribution to dollar inflation is invalid. Further research needs to address the non-profit organization in particular and the above issues specifically.

Finally, the research presented here has attempted to develop the concept of a Delta Base Rate theory and application. Although the results were not as dramatic as anticipated, nonetheless, the findings provide some evidence that the principles underlying the concept were demonstrated.

Delta Base Rate was shown to increase as a function of additional hurdles as well as being inversely related to changes in the hiring rate.

The /\HR concept in the multiple hurdle selection process, in effect, reduces the artifically high selection ratio found in this method of selection. Difficulty in showing improvement over an extreme base rate was also demonstrated.

Evidence now exists that individual hurdles in the multiple hurdle selection process can be evaluated singularly as well as the determination of their value within the context of the total selection process. Further research should include continuation of the present method to develop a process whereby the optimum number of hurdles and their placement can be determined. One method would be to place potential hurdles at different points in the selection process to derive the greatest benefit and to determine whether the addition of another hurdle is cost-effective for the resultant gain in prediction. The present method also provides a means of combining certain hurdles, or the inclusion of an additional hurdle simultaneously, to one

currently used to achieve the best prediction.

In conclusion, the utility estimation methods to date lack the complete detail needed for application to the non-profit and intangible performance criterion situation. Adaptation of the current formulas and using training as an estimator of the ultimate criterion has partially evaluated this unique selection environment. In viewing the results of utility analysis from both the standpoint of variable selection parameters and the Delta Base Rate estimations for reducing the number of people entered into training, the most beneficial representation of the usefulness of a predictor in this instance can be made.

BIBLIOGRAPHY

Air Force Association, (1984). An Air Force almanac, Air Force Magazine, May, 175-179.

Air Force Association, (1985). An Air Force almanac, Air Force Magazine, May, 189-193.

Air Force Association, (1986). An Air Force almanac, Air Force Magazine, May, 181-185.

Air Force Association, (1987). An Air Force almanac, Air Force Magazine, May, 79-83.

Albright, L.E., & Glennon, J.R. (1961). personal history correlates of physical scientists' career aspirations. Journal of Applied Psychology, 45, 281-284.

Alexander, R.A., Barrett, G.V., & Doverspike, D. (1983). An explication of the selection ratio and its relationship to hiring rate. Journal of Applied Psychology, 68(2), 342-344.

Alexander, R.A., & Barrick, M.R. (1987). Estimating the standard error of projected dollar gains in utility analysis. Journal of Applied Psychology, 72(3), 475-479.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. Educational and Psychological Measurement, 10, 67-78.

Anastasi, A., Meade, M.J., & Schneiders, A.A. (1960). The validation of a biographical inventory as a predictor of college success. Educational Testing Service, V, 81 pp.

Arnold, J.D., Rauschenberger, J.M., Soubel, W., & Guion, R.M. (1982). Validation and utility of a strength test for selecting steelworkers. Journal of Applied Psychology, 67, 588-604.

Bordelon, W.P. & Kantor, J.E. (1986). Utilization of psychomotor screening for USAF pilot candidates: Independent and integrated selection methodologies. AFHRL-TR-86-4, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

Bortner, D.E., & Ree, M.J. (1977). Cost analysis of pilot selection systems. AFHRL-TR-77-55. Air Force Human Resources Laboratory, Brooks AFB, TX.

Boudreau, J.W. (1989). Selection utility analysis: A review and agenda for future research in Advances in Selection and Assessment, Smith, M. & Robertson, I.T. (eds). New York: Wiley.

Brogden, H.E. (1946a). On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 37, 65-76.

Brodgen, H.E. (1946b). An approach to the problem of differential prediction. Psychometrika, 11,, 139-154.

Brogden, H.E. (1949). When testing pays off. Personnel Psychology, 2, 171-183.

Brogden, H.E. & Taylor, E.K. (1950). The dollar criterion: Applying the cost accounting concept to criterion construction. Personnel Psychology, 3, 133-154.

Buel, W.D. (1965). Biographical data and the identification of creative research personnel. Journal of Applied Psychology, 49, 318-321.

Campbell, J.P. (1983). Psychometric Theory. In Dunnette, M.D. (Ed.) Handbook of Industrial Psychology. New York: Wiley & Sons.

Cascio, W.F. (1982). Costing Human Resources: The Financial Impact of Behavior in Organizations. Boston: Kent.

Cronbach, L.J. & Gleser, G.C. (1957). Psychological Tests and Personnel Decisions. Urbana, IL: University of Illinois Press.

Cronbach, L.J. & Gleser, G.C. (1965). Psychological Tests and Personnel Decisions (2nd Ed.). Urbana, IL: University of Illinois Press.

Curtis, E.W., & Alf, E.F. (1969). Validity, predictive efficiency, and practical significance of selection tests. Journal of Applied Psychology, 53,, 327-337.

Greer, D.L. & Cascio, W. (1987). Is cost accounting the answer: Comparison of two behaviorally based methods for estimating the standard deviation of job performance in dollars. Journal of Applied Psychology, 72(4), 588-595.

Guion, R. (1965). Personnel Testing. New York: McGraw-Hill.

Hays, W.L. (1981). Statistics (3rd Ed.). New York: Holt, Rinehart and Winston.

Helmreich, R., Bakeman, R., & Radloff, R. (1973). The life history questionnaire as a predictor of performance in Navy diver training. Journal of Applied Psychology, 57, 148-153.

Hogan, J. & Zenke, L.L. (1986). Dollar-value utility of alternative procedures for selecting school principals. Educational and Psychological Measurement, 46, 935-945.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance, Psychological Bulletin, 96(1), 72-98.

Kantor, J.E. & Bordelon, V.P. (1985). The USAF pilot selection and classification research program, Aviation, Space, and Environmental Medicine, 56, 258-261.

Kantor, J.E. & Carretta, T.R. (1988). Aircrew selection systems, Aviation, Space, and Environmental Medicine, 59(11, Suppl.), A32-38.

Katzell, R.A. & Guzzo, R.A. (1983). Psychological approaches to productivity improvement. American Psychologist, April, 468-472.

Kulberg, G.E., & Owens, W.A. (1960). Some life history antecedents of engineering interests. Journal of Educational Psychology, 51, 26-31.

Landy, F.J. (1986). "Utility Theory: Work to be done". Presented at the First Annual Conference of the Society for Industrial and Organizational Psychology Inc. Utility Analysis for Practitioners: A Workshop. Chicago, IL.

Laurent, H. (1962). Early identification of management talent. Management Record, 24(5), 33-38.

McCormick, E.J. & Tiffin, J. (1974). Industrial Psychology (6th Edition). Englewood Cliffs, NJ: Prentice-Hall.

Miller, R.E. (1966). Relationship of AFOQT scores to measures of success in undergraduate pilot and navigator training. PRL-TR-66-14, Personnel Research Laboratory, Lackland Air Force Base, Texas.

Muchinsky, P.M. (1983). Psychology Applied to Work. Homewood, IL: Dorsey.

Mumford, M.D. & Owens, W.A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. Applied Psychological Measurement, 11(1), 1-31.

Owens, W.A. (1983). Background Data. In Dunnette, M.D. (Ed.) Handbook of Industrial Psychology. New York: Wiley & Sons.

Rawls, D., & Rawls, J.R. (1968). Personality characteristics and personal history data of successful and less successful executives. Psychological Reports, 23, 1031-1034.

Roach, B.W. (1983). Monetary value of pilot selection using the AFOQT. Proceedings of the 25th Annual Military Testing Association Conference (pp.478-483). Gulf Shores, AL.

Rogers, D.L., Roach, B.W., & Short, L.O. (1986). Mental ability testing in selection of Air Force officers: A brief historical overview. AFHRL-TP-86-23, Air Force Human Resources Laboratory, Brooks AFB, TX.

Roomsburg, J.D. (1988). Biographical Data as Predictors of Success in Military Aviation Training. Unpublished Master's Thesis, The University of Texas, Austin, TX.

Schmidt, F.L. (1974). Probability and utility assumptions underlying use of the Strong Vocational Interest Blank. Journal of Applied Psychology, 59(4), 456-464.

Schmidt, F.L. & Hunter, J.E. (1981). Employment testing: Old theories and new research findings, American Psychologist, 36, 1128-1137.

Schmidt, F.L. & Hunter, J.E. (1983). Individual differences in productivity: An empirical test of estimation derived from studies of selection productivity utility, Journal of Applied Psychology, 68(3), 407-414.

Schmidt, F.L., Hunter, J.E., & Dunn, W.L. (1987). Potential utility increase from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB). Navy Personnel Research and Development Center, San Diego, CA 92152-6800.

Schmidt, F.L., Hunter, J.E., McKenzie, R.C., & Muldrow, T.W. (1979). Impact of valid selection procedures on work-force productivity. Journal of Applied Psychology, 64(6), 609-626.

Schmidt, F.L., Hunter, J.E., Outerbridge, A.M., & Trattner, M.H. (1986). The economic impact of job selection methods on the size, productivity and payroll costs of the Federal workforce: An empirical demonstration. Personnel Psychology, 39,, 1-29.

Schmidt, F.L., Mack, M.J., & Hunter, J.E. (1984). Selection utility in the occupation of U.S. Park Ranger for three modes of test use. Journal of Applied Psychology, 69(3), 490-497.

Schmidt, F.L. & Rauschenberger, J.M. (1986). Utility Analysis for Practitioners: A Workshop. Presented at the First Annual Conference of the Society for Industrial and Organizational Psychology, Inc. Chicago, IL.

Siegel, L. & Lane, I.M. (1987). Personnel and Organizational Psychology (2nd Edition.). Homewood, IL: Irwin.

Skinner, J., & Ree, M.J. (1987). Air Force officer qualifying test (AFOQT): Item and factor analysis of Form O. AFHRL-TR-86-68. Air Force Human Resources Laboratory, Brooks AFB, TX.

Smith, M. (1948). Cautions concerning Taylor & Russell. Journal of Applied Psychology, 32, 595-500.

Stoker, P., Hunter, D.R., Kantor, J.E., Quebe, J.C., & Siem, F.M. (1987). Flight screening program effects on attrition in undergraduate pilot training. AFHRL-TP-86-59, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

Taylor, H.C. & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 23, 565-578.

Thorndike, R. (1949). Personnel Selection. New York: Wiley.